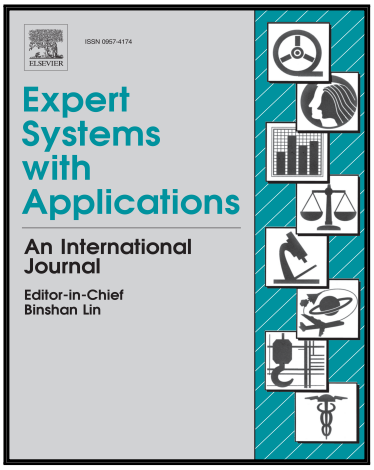
Journal Pre-proof



Fighting Post-truth using Natural Language Processing: A review and open challenges

Estela Saquete, David Tomás, Paloma Moreda,
Patricio Martínez-Barco, Manuel Palomar

PII: S0957-4174(19)30661-X
DOI: https://doi.org/10.1016/j.eswa.2019.112943
Reference: ESWA 112943

To appear in: *Expert Systems With Applications*

Received date: 1 May 2019
Revised date: 24 July 2019
Accepted date: 7 September 2019

Please cite this article as: Estela Saquete, David Tomás, Paloma Moreda, Patricio Martínez-Barco, Manuel Palomar, Fighting Post-truth using Natural Language Processing: A review and open challenges, *Expert Systems With Applications* (2019), doi: https://doi.org/10.1016/j.eswa.2019.112943

Highlights

- The study describes the problem of fake news phenomena in digital information
- The study provides a systematic review of the state of the art regarding automatic fake news detection
- From the review, main subtasks involved in automatic fake news detection are detected and classify
- The review covers systems, resources and competitions in automatic fake news detection
- The review outlines gaps in knowledge and future challenges related to automatic fake news detection

Fighting Post-truth using Natural Language Processing: A review and open challenges

Estela Saquete, David Tomás, Paloma Moreda, Patricio Martínez-Barco, Manuel Palomar

Department of Software and Computing Systems, University of Alicante
Apdo. de Correos 99 E-03080, Alicante, Spain
{*stela,dtomas,moreda,patricio,mpalomar*}*@dlsi.ua.es*

Abstract

Post-truth is a term that describes a distorting phenomenon that aims to manipulate public opinion and behaviour. One of its key engines is the spread of Fake News. Nowadays most news is rapidly disseminated in written language via digital media and social networks. Therefore, to detect fake news it is becoming increasingly necessary to apply Artificial Intelligence (AI) and, more specifically Natural Language Processing (NLP). This paper presents a review of the application of AI to the complex task of automatically detecting fake news. The review begins with a definition and classification of fake news. Considering the complexity of the fake news detection task, a divide-and-conquer methodology was applied to identify a series of subtasks to tackle the problem from a computational perspective. As a result, the following subtasks were identified: deception detection; stance detection; controversy and polarization; automated fact checking; clickbait detection; and, credibility scores. From each subtask, a PRISMA compliant systematic review of the main studies was undertaken, searching Google Scholar. The various approaches and technologies are surveyed, as well as the resources and competitions that have been involved in resolving the different subtasks. The review concludes with a roadmap for addressing the future challenges that have emerged from the analysis of the state of the art, providing a rich source of potential work for the research community going forward.

Keywords: Natural Language Processing, Fake News, Post-truth, Deception Detection, Automatic Fact-checking, Clickbait detection, Stance Detection, Credibility, Human Language Technologies, Applied computing,

Document management and text processing, Document capture, Document analysis

1. Introduction

The "post-truth" term refers to a distorting phenomenon where objective facts are less influential in shaping public opinion than appeals to emotion and personal belief.[1] In fact, post-truth was originally used by political commentators, but nowadays the term has a much wider application in the news generation process. In a post-truth era alternative facts replace actual facts, and feelings have more weight than evidence (McIntyre, 2018).

One of the key engines of the post-truth era is the proliferation of fake news, which has been facilitated by the growth of digital and social media personal blogs, alternative media and social networks including Twitter, Facebook, WhatsApp (Vosoughi et al., 2018).

In the post-truth era anyone can be a potential journalist and fact checking is less of a priority than sharing news that could potentially go viral. Fake news is a relatively new term, and it was defined by The New York Times as a "made up story with the intention to deceive, often with monetary gain as a motive"(Tavernisen, 2019). This type of news creates great confusion on a viral scale concerning the real facts with the aim of forming or manipulating public opinion so as to influence socio-political behaviour/mass belief systems. Indeed, information bubbles and echo chambers are derived consequences of fake news and they may prevent existing perspectives to be challenged. The ideological and economic interests that potentially gain from this "information disorder" are what drive fake news:

- Ideological interests: Aim to manipulate social opinion, reinforce preconceived opinions so as to focus people on thinking or acting in a specific way. For instance, this distorting phenomenon played an important role in *President Trump's election campaign 2016* (Bovet & Makse, 2019) and *the Brexit referendum 2016* (Bastos & Mercea, 2019).

- Economic interests: Money can be made, through clickbaits and misleading information, by individuals and companies that fabricate fake news. Many companies generate huge profits creating fake news, such

[1] Chosen as the 2016 year word by Oxford dictionary

as National Report website, or Disinformedia (Hooper, 2018) or Victory Lab (Issenberg, 2013). Besides, creating fake news may be cheaper, given that fact checking is avoided.

Contribution of this paper: The complexity of fake news detection – such as, volume; speed; and propagation– has made it necessary to rely on automatic processes to tackle the problem. The objective of this paper is to provide a comprehensive and systematic review of the state of the art related to the main tasks involved in automatic fake news detection, including systems, resources and competitions. This review has been performed using a robust methodology to minimize bias in the gathering, summarizing and presenting of research evidence. An important element of this review involves identifying gaps in knowledge which can be used to guide future research efforts. It aims to provide researchers with a valuable reference from which future challenges related to fake news detection can be addressed.

Organization of this paper: The paper is organized as follows. Section 2 describes the methodology applied in the systematic review. Section 3 contains a detailed description of the fake news detection problem, especially regarding written content in different media (news and social networks). The different subtasks, that emerge when tackling the problem from a computational angle, are also classified. Section 4 is an exhaustive review of the state of the art in relation to the classified subtasks. Section 5 deals with open issues in fake news detection. Section 6 presents the conclusions.

2. Review methodology

A systematic review of studies based on PRISMA[2] guidelines was undertaken to analyze the literature assessing the problem of fake news detection from a Natural Language Processing (NLP) and Artificial Intelligence (AI) perspective. Some adaptations to PRISMA guidelines were necessary due to the peculiarities of the topic.

Specifically, the free tool Harzing's Publish or Perish[3] was used, which allows a systematic search of different databases, their instant classification by extracting the most relevant data, and the calculation of metrics. This

[2]http://prisma-statement.org/PRISMAStatement/ (accessed online 28 February, 2019)

[3]https://harzing.com/blog/2017/11/publish-or-perish-version-6 (accessed online 28 Feb., 2019)

tool will serve as a basis for the application of filters and selection of papers to be included in the study.

The review protocol consists of five steps that indicate how the review has been conducted and reported. The protocol considers that the fake news detection task is multidisciplinary, and focuses on NLP and AI approaches.

Step 0: "Divide-and-conquer". In this step, an initial search of the task's generic terms has been carried out to determine the most relevant research work. The protocol to establish generic keywords consists of indicating the terms of the task along with the most discriminating terms, allowing for the exclusion of studies on the requested topic which focus on other research areas. Later, in section 3.1., the specific generic terms applied in this step are reported. From this, a meaningful enough number of results will be selected, which is determined by the repetition of terms already found and by the lack of a significant new contribution of terms. This number will depend on the task in hand. For instance, in our research this number was set to 100 since beyond that number the terms became repetitive with no new ones being added. From this point, the most frequent terms were extracted, and a subtask classification was derived by manually aggregating keywords. This classification led us to a more exhaustive search for each one of those tasks. Once the subtasks and their representative terms are identified, the following steps are performed for each one.

Step 1. "Search and Identification". A specific search for each detected subtask is performed. The search uses the set of keywords related to the subtask as well as the range of dates considered relevant for each of the tasks, and launches a Google Scholar search using Harzing's Publish or Perish tool.

Step 2. "Screening: Coarse grain filter". This step performs a filtering according to metadata value. The results are sorted by number of citations and those papers that contribute to the h-index are selected. From this initial list, the following are eliminated: a) papers whose sources ("Publisher") are not among the databases selected as relevant for the topic: ACM (dl.acm.org), ACL (aclweb.org), Arxiv (arxiv.org), IEEE Xplore (ieeexplore.ieee.org), Elsevier, MIT Press, Wiley Online Library; b) papers whose "Publication" belongs to a subject clearly outside the scope of NLP; c) papers whose "Title" clearly invokes a subject matter outside the scope of the study, or indicates a survey or review; d) papers whose "Type" is not of interest (BOOK, HTML, CITATION, PATENT,...). Finally, the resulting h-index is recalculated.

Step 3. "Eligibility: Fine grain filter". This step performs a filtering according to paper content. The abstracts of previous papers are checked and those most relevant to the task are selected based on any of the following criteria: NLP related topic; comparable systems; participation in a competition specific to the task; and, original or recently published approach.

Step 4. "Other papers included". This step consists of adding to the review those papers that are missing in steps 2 and 3 of the protocol, but are cited by or related to those selected in step 3. If, after in-depth study, they are considered relevant and/or are an important basis for the research in which they are cited, they are included in the systematic review of the task in question.

3. Fake News Detection

The problem of fake news dissemination is considered endemic and several organizations are making headway in the fight against it. First Draft,[4] for example, grew out of a collaboration between nine founding organizations in June 2015 to raise awareness, research and address challenges relating to trust and truth in media in the digital age. First Draft News coined the term "information disorder" to refer to fake news and it has published a detection guide to help journalists and researchers[5].

The fake news phenomenon tends to comprise transversely the following features and the greater their degree of presence, the greater the probability of the news being fake: i)*Impact*- The more false information impacts, the better. In this sense the emotional triggers help to increase the impact; ii)*Scarcity*- Data is an important part of news in general, so for a false story to be credible it will be necessary for the news story to look real but there must be a scarcity of cross-checkable information, making it impossible to readily dismantle the news story after verification; iii)*Relevant topic*- All fake news is driven by an interest, ideological or economic, as remarked in the introduction. That is why the topic of the news must be socially relevant; and iv)*Viralization*- False news aim to maximize audience reach, with rapid

[4]https://firstdraftnews.org/ (accessed online 28 February, 2019)

[5]Report of the Conference Combating Fake News is worthy of consideration: https://shorensteincenter.org/wp-content/uploads/2017/05/Combating-Fake-News-Agenda-for-Research-1.pdf (accessed online 19 July, 2019)

proliferation being another important feature of this type of news (Amoros, 2018).

Once the problem has been contextualized in general, and considering that digital information is disseminated exponentially, artificial intelligence plays a fundamental role, and more specifically natural language processing and machine learning approaches (Dale, 2017). Additionally, from a computational perspective, depending on whether the problem emerges in digital traditional media platforms or in social media networks, it is interesting to consider the following source classification and which are the features and approaches applied for each one (Shu et al., 2017):

- Digital traditional media
  - *Knowledge-based*: external sources fact check the truthfulness of the claims in news content.
  - *Style-based*: Fake news publishers often have malicious intent to spread distorted and misleading information aimed at mass markets, using appealing and persuasive writing styles that are not used in real news articles. Style-based approaches aim to detect fake news by identifying potentially manipulative writing styles.
- Social media
  - *Stance-based*: These approaches utilize users viewpoints/reactions, extracted from the content of relevant posts to infer the veracity of original news articles.
  - *Propagation-based*: These approaches examine the relationship between relevant social media posts to build a credibility model which propagates credibility values between users, posts, and news. The veracity score of a news piece is an aggregation of the credibility values of each relevant social media post.

3.1. Tasks in Fake News Detection

Since assessing the veracity of a news story is complex from an engineering point of view, the research community is approaching this task from different perspectives. This fact was corroborated after applying Step 0 of the defined review protocol presented in section 2. In this step, the search (removing duplicates) for the generic keywords was launched, with the term related to

the task being "fake news" plus the discriminator terms being tasks and NLP. The frequent terms that emerge in the first 100 papers have been selected. The discovery of this cloud of keywords related to the research question by manually aggregating them, resulted in a subtask classification of the different subtasks in the existing research and their related competitions. Therefore, similar approaches and resources to solving the same subproblem are grouped together. Issues like satire or humorous news articles are beyond the scope of this review.

The subtasks detected from the general research question are explained next. Since the fake news detection problem is tackled in the literature from different angles, the different subtasks are focused on one or more fake news features. Next, for each subtask, the existing fake news features are identified and presented. In addition, regarding digital source classification (traditional media/social media), as extensively discussed in section 4, some subtasks address each type of source with different approaches, while others do it jointly. This is briefly indicated in the definition of the subtasks presented below:

- **Deception Detection**: Detecting deception in communications has been a challenge throughout history. However, since the early 20th century several technologies have been developed that are specifically aimed at unmasking deception, primarily through the identification and analysis of cues possibly associated with false statements. The cues have varied widely, ranging from physiological measurements to non-verbal and verbal behaviors. Therefore, the main feature being tackled in this subtask is dealing with *scarcity* of information by means of looking for linguistic cues. Regarding the source, the problem of deception detection is addressed as a linguistic problem independently of the source to be targeted.

- **Stance Detection, Controversy and Polarization**: Does the heading represent news content? What is the stance of the audience over a topic/issue? Is it possible to quantify controversy in social media? Are people's point of view reinforced by equals and different points of view are rejected and keeping out of my own echo chamber? Stance detection concerns the use of AI and NLP technologies to detect stances behind a topic. This is why relevant *topics* are the main feature of this subtask. As presented in-depth in section 4, the approaches regarding stance detection in the literature are different in traditional and social

media. In the case of polarity task, *impact* is the main feature present, since emotion triggers need to be considered to detect it. Controversy then uses not only the *impact* feature of polarization but also takes into account the *viralization* of information. Controversy and polarization in the literature focused on using social media data sources.

- **Automated Fact checking**: Is the claim true, false or deceptive? Automated fact checking consists of automatically checking the veracity of a public claim against all the available data, and classifying it into one of the following veracity values commonly used by human fact checkers: True, Mostly True, Half True, Mostly False and False. In this case, features related to claims on *relevant topics* and with an indication of information *scarcity* are often present. However, considering *impact* and *viralization* is also important when contextual references are used to tackle the subtask, as explained in-depth in section 4.3. The literature on automated fact checking is dealing jointly with different sources.

- **Clickbait Detection**: When is a headline a "clickbait"? Clickbait is the term applied to the content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page, but very commonly its content does not necessarily have to do with the headline. The clickbait detection task deals with *impact, viralization* and a *relevant topic*. In the literature, the problem is addressed in the same way both for traditional and social media.

- **Credibility**: Is it possible to measure credibility of online information? Are media and sources reliable? Credibility refers to trustworthiness in terms of the media, the information source and the message. Therefore, the main features being assessed in the credibility task are *impact, scarcity* and *viralization*. For this subtask, the literature makes a distinction between approaches used for traditional media and those for social media.

In Figure 1, the detected subtasks are presented, as well as a proposal of interconnection to each other.

After a detailed study of the literature extracted in the systematic review, a multi-level abstraction-based structure is created which establishes a set of

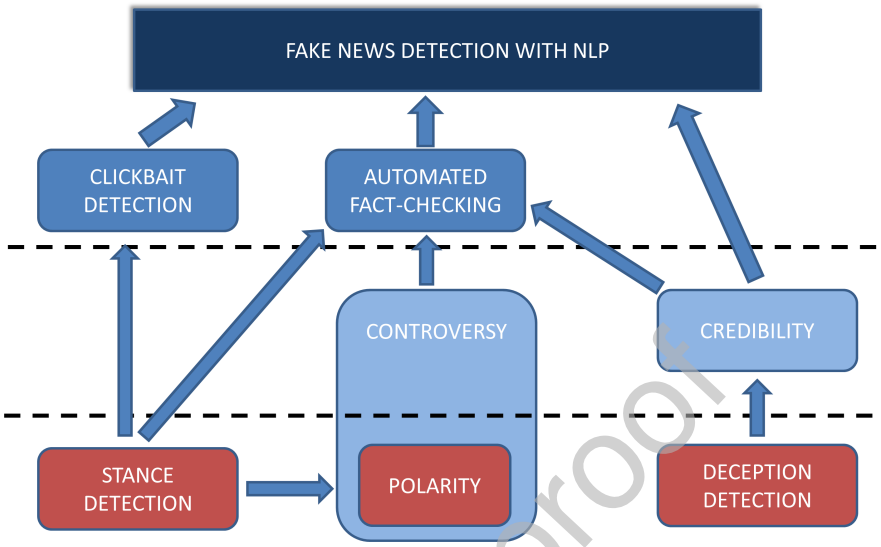


Figure 1: Interrelationship and abstraction levels of the subtasks identified by the systematic review of Fake News Detection using NLP (Source: Own study)

relations between the detected subtasks based on the type of task in question and the specific features that affect it. In this way, the most primitive subtasks appear in the lower levels, and as levels become more sophisticated, the derived subtasks, which appear to feed on their prior primitive ones, are applied to more specific cases. This structure is visualized in Figure 1, showing the most basic subtasks at the bottom level of the figure, and the derived ones at upper levels. The interrelationship between them is indicated by the directional arrows.

It is important to remark that although all the subtasks represented in Figure 1 deal with issues related to fake news detection using NLP at some point, depending on the specific problem to be solved, a real connection between them may not always exist, so that the structure presented here is in fact a conceptual connection between the subtasks. Each one contributes to solve part of the problem; however, how to effectively combine the technologies to achieve a global fake news detection system is still an open challenge.

The most basic subtasks are *deception detection*, *stance detection* and *polarity*. The outcomes of these subtasks are used in derived ones such as: *credibility*, which relies on deception detection in order to be measured; po-

larity, which can be seen as a simplified version of *controversy*, and relies on stance detection; or, *clickbait detection*, which is a task in part supported by stance detection. In the same way, *fact checking* uses: stance detection to determine the veracity of a claim that has already been checked before but with other words or expressions; credible knowledge bases to be able to determine the veracity of the claim; and, the controversy of social media to determine the category of a claim.

Hence, clickbait detection, automated fact checking and credibility are the three derived subtasks that encompass, albeit independently, the current research in fake news detection that uses NLP. In addition, determining how the result of the different subtasks affects the determination of whether a text contains false information is also at this time an unresolved task. Even so, thanks to this systematic review, open issues have been identified, which could be investigated in the future and easily incorporated into Figure 1, which is an evolving conceptual map.

The relationship between subtasks is also established through fake news features. For instance, clickbait detection is inheriting the relevant topic from stance detection and tackling impact and viralization. In the case of credibility, the scarcity of information is inherited from deception detection task, but also impact and viralization is used. Controversy considers not only the impact assessed in the polarity subtask but also viralization of the messages. Finally, automated fact checking deals with all the features coming from the subtasks on which it conceptually depends.

In Table 1 a summary of the review is presented, indicating the main fake news features for each subtask, the number of resources reported and the number of systems presented. When the subtask reports different approaches for traditional media and social media, the number of systems is presented for each of the sources. The next section includes a description of each subtask, with a complete definition of the task and their specific characteristics, as well as all the available resources and competitions regarding the task, the main systems found in the literature and a discussion that compares systems and detects gaps in the subtask research.

4. Existing work

This section provides an in-depth outline of the literature related to Fake News Detection using NLP and AI. As previously mentioned, this task is highly complex and for this reason, a classification of the main subtasks was

Subtask	FN features	No. Resources	No. Systems
Deception Detection	Scarcity	19	11 (9 ML/2 DL)
Stance Detection	Relevant topic	5	Traditional media: 9 (3 ML/5 DL/ 1 Hybrid) Social Media: 4 (2 ML/2 DL)
Polarity ***Controversy***	Impact Viralization	1	4 (2 ML/2 Graph models)
Fact checking	Relevant topic Scarcity Impact Viralization	8	13 (References: 5/KG: 4/Context: 4)
Clickbait Detection	Relevant topic Impact Viralization	5	9 (5 ML/3 DL/1 Hybrid)
Credibility	Impact Scarcity Viralization	31	Traditional media: 4 Blogs: 3 Social media:11 (18 ML)

Table 1: Summary of the main Fake News features for each subtask as well as the number of resources and systems reported in the review, where ML=Machine Learning, DL=Deep Learning and KG=Knowledge Graphs.

derived from the general research question. To select the most relevant papers for each subtask, a set of terms was used as search keywords. As interest in this research topic is relatively recent, the search was restricted to the last five years (from 2014 to January 2019). The search was conducted in Google Scholar, using Harzings Publish or Perish tool, and it obtained a number of papers in Step 1. After completing the Step 2 filter (screening), those papers that contribute to the h-index and whose subject matter or typology is of interest were selected. Finally, after applying Steps 3 and 4 of the methodology (eligibility and other papers included), a set of relevant studies was obtained. Table 2 presents the keywords and the number of articles selected for each investigated subtask and each step of the methodology.

4.1. Deception Detection

Deception is information intentionally transmitted to create a false impression or conclusion (Burgoon et al., 1996). As the popularity of virtual media has grown, people's dependence on digital media has also increased for interpersonal communication, information acquisition and dissemination of information (Zhou & Zhang, 2008). According to Rubin et al. (2015), we are in the "Digital deception" period. The term signifies deception in the context

Subtask	Keywords	Step 1	Step 2	Step 3-4
Deception Detection	"deception detection" "automatic"	995	19	33
Stance Detection	"stance detection"	654	21	19
Controversy/Polarization	"controversy detection"	125	11	4
Fact checking	"content veracity" "fact check" "fact checking" "fact verification" "fake news detection" "fact news identification" "rumour detection" "rumour veracity"	979	25	36
Clickbait detection	"clickbait" "fake news headline classification" "misleading detection"	997	10	10
Credibility	"credibility" "news" "blog" "social newtwork"	213	62	24

Table 2: Summary of keywords and number of articles derived from applying the review methodology for each subtask

of information and communication technology and redefines deception in the digital context as an intentional control of information in a technologically mediated environment to create a false belief or false conclusion. This virtual environment provides fertile ground for deception (Zhou & Zhang, 2008).

Interest in detecting deception originated in the fields of philosophy, psychology, sociology, criminology and anthropology (Mihalcea & Strapparava, 2009). The studies developed in these areas were tackled manually. These early studies focused on deception detection regarding face-to-face communication and their mental (the information must be believable and consistent), emotional (guilt or shame) and physical (eye movement or arm positions) levels (Vartapetiance & Gillman, 2012). The results obtained proved that the accuracy of human deception detection is around 54%, a little better than chance (George et al., 2016). These results showed the need to detect deception by automatic methods. The huge amount of information currently available on the Web makes manual processing impossible (Zhou et al., 2004b). Moreover, since 9/11, automatically detecting digital deception, as well as anticipating and preventing terrorist attacks became a challenging priority (Burgoon et al., 2003).

The type of data most frequently encountered on digital information is written language (Mihalcea & Strapparava, 2009) and researchers have fo-

cused their attention on how to detect digital deception only in written language. However, by focusing exclusively on written language in digital media, the deception detection task is more complex compared to traditional deception detection techniques that used non-verbal cues(Burgoon et al., 2003).

Recently, advances in NLP have developed deception detection systems that show a certain degree of capability in terms of discriminating between truth and falsehood in digital texts. Regardless of the specific approach used, the proposed systems generally comprise four phases (Zhou & Zhang, 2008): (1) identifying and extracting deception cues; (2) building a deception detection model using the identified cues; (3) applying the deception model; and (4) making a detection decision. Considering these four phases, information about cues is presented in section 4.1.1, and the resources being used to obtain these cues in section 4.1.2. The systems and approaches proposed in the scientific literature are analyzed in section 4.1.3. Finally, section 4.1.4, the discussion section, presents conclusions on the state-of-the-art regarding deception detection subtask.

4.1.1. Cues

Deception involves the manipulation of language and the assiduous construction of messages or stories so that they appear truthful to others. Although there is no sign of deception itself, there are some specific cues present in these texts, indicating observable differences between deceptive and truthful texts (Zhou & Zhang, 2008). This is the reason why the identification of deception cues is the first step in automating deception detection (Zhou et al., 2004a).

A wide range of deception cues have been developed from traditional deception research in psychology and criminal investigation practice. DePaulo et al. (2003) developed a list of 158 visual, verbal and vocal deception cues extracted from an analysis of 116 research papers between 1920 and 2001. For example, compared with truth, deception contains superfluous repetitions of words or phrases, and incomplete sentences, as well as fewer unique words and self-references. Deception also shows more negative emotion, and sounds more evasive, unclear, uncertain, and impersonal (Zhou & Zhang, 2008).

In general, the cues can be classified as *nonverbal* (arm position or eye movement) and *verbal* (choice of words). One or several cues may be involved in a single communication, but some will be more specific to certain types of communication. For example, body language and eye movements are

mainly considered in synchronous, non-distributed communication, while the structure of the sentence will be more obvious in distributed communication such as instant messages or emails (Vartapetiance & Gillman, 2012).

This work focuses on written language, making the text the only source for verifying credibility (Zhou et al., 2004b). The cues based on general linguistic knowledge such as self-reference, meaning of words or the ratio of syllables to words, are called *linguistic-based cues* or *linguistic cues* (Zhou et al., 2004a).

Moreover, the online language differs from that used in traditional communication. But even more, the characteristics of different media should be taken into consideration. Hence, different types of online communication result in different linguistic cues to deception (Zhou & Zhang, 2008).

Therefore, although traditional linguistic deception cues rooted in the field of psychology are a good starting point for studying deception in online communications, the online deception scenario required the consideration of new linguistic cues (Zhou & Zhang, 2008). Otherwise, inconsistencies, contradictions and other difficulties would be encountered as Vartapetiance & Gillman (2012) showed.

Finding cues that indicate deception via manual inspection is complex. This is why researchers are adopting NLP approaches (morphological, syntactic, lexical or semantic parsing) as well as statistical and machine learning methods, and different lexical resources (Newman et al., 2003). They have enabled linguistic-based cues (LBC) to be automatically extracted and analyzed from texts. This field is known as Automatic Linguistics-Based Cues Detection.

NLP tools were used by Burgoon et al. (2003) to perform tests with sixteen linguistic cues that can be automated to return assessments of the likely truthfulness or deceptiveness of a piece of text. Their study, carried out on text and audio chat, concluded that deceiver messages were briefer (i.e., lower on quantity of language), were less complex in their choice of vocabulary and sentence structure, and lacked specificity or expressiveness in their text-based chats. However, the authors highlight the existence of differences between synchronous and asynchronous experiments. As a conclusion, they state that different cue models will be required for the different tasks. Studies such as Zhou et al. (2004a,b) extracted twenty-seven LBC (grouped in terms of quantity, complexity, uncertainty, non-immediacy, expressivity, diversity and informality). After a systematic analysis using a set of measures, they finally concluded that all the linguistic features they considered

are potentially relevant discriminators in the context of text based computer mediated communication. However, the same cue profiles are unlikely to apply uniformly across contexts. Subsequently, the same authors (Zhou & Zhang, 2008) showed a summary of linguistic cues for online texts, regardless of the context. Table 3 contains a summary of the main linguistic cues found in the different studies. The authors similarly conclude that some LBCs are effective for both online communication -synchronous and asynchronous- whereas other cues are identified only in one of the modes of communication. Thus, the LBCs identified in one deception context may not be applicable to another one.

Linguistic cues	**Behavior in deception**
Quantity: the amount of information	more words and sentences used
Word diversity	lower lexical and content diversity
Redundancy	superfluous repetitions of words or phrases
Language complexity	less complexity in sentence length, average word length, incomplete sentences, unstructured texts, fewer unique words
Expressiveness: frequency of using adjectives and adverbs	greater expressiveness
Non-immediacy: expression to disassociate oneself from his/her message content	fewer self-references or group references, more modal verbs, present tense used instead of past or future, more passive voice, more objectification, more generalizing terms
Informality	more informal texts
Cognitive complexity: the ratio of cognitive operations	higher cognitive complexity, syntactically complex expressions
Affect : positive or negative emotions	more positive and especially negative affect present
Spontaneous correction: ratio of immediately corrected messages	lower spontaneity
Uncertainty: verbal uncertainty indicating the lack of sureness	more uncertain, more evasive, more unclear, more impersonal, more discrepant, more ambivalent
Non-Contextual embeddings	few spatio-temporal information, few details in the message regarding actions, description of people, events

Table 3: Summary of the most used linguistic cues in the review of studies dealing with Deception Detection (Zhou & Zhang, 2008)

Finally, although it seems logical to think that linguistic cues are the most significant for deception detection in texts, some authors, such as Conroy et al. (2015), proposed studying the contribution of other types of cues, such as network or behavior, in combination with LBCs. They argue in favor of developing hybrid approaches based on both content properties and media computer communication patterns, for instance, learned patterns of

argumentation style classes. So, language patterns could be complemented with message metadata or knowledge networks.

4.1.2. Resources

Data collection is one of the challenges of conducting deception research due to the scarce availability of such datasets. As with all others NLP tasks, creating a corpus can be done manually or automatically.

Regarding *manually annotated corpora*, several resources have been developed for different domains and languages. The first studies in this line were carried out by means of controlled experiments, either with sanctioned or unsanctioned approaches. A common example of a sanctioned controlled experiment is recruiting participants for a study on deception and randomly assigning them to a lie or truth condition, such as in Newman et al. (2003). This work compiles 568 videos and texts in English about abortion and feelings about friends. RusPersonality (Litvinova et al., 2016) is a Russian corpus for authorship profiling that can be used for deception detection. The corpus contains over 1,850 documents with an average length of 230 words. The annotation process involved the participation of 114 respondents. The first step involved theme describing the events of a day in their life, and the second step involved them in consciously changing the facts. Almela et al. (2012) collected a Spanish dataset of 600 statements for three topics (opinions about homosexual adoption, bullfighting and feelings about best friends). For each topic, 200 statements were generated, 100 true and 100 false, with an average of 80 words per statement. Manual verification of the quality of the contributions was made.

The main advantage of this type of experiment is that researchers have more control. However, the main limitation is that the researcher is giving permission to the participant to lie and this could affect their behavior (Gokhman et al., 2012). In unsanctioned approaches, such as that of DePaulo et al. (2003), the participant lies of his or her own accord. In this work, messages in English of 120 independent samples were generated where senders described their attitudes or personal facts, films, slides, pictures or transgressions.

Fitzpatrick & Bachenko (2012) propose a set of guidelines for building corpora with the aim of testing deceptive models to avoid the problem of sanctioned lying that is typically required in a controlled experiment. They have extensive experience in obtaining data from court cases and other testimonies, and uncovering the background information that enabled them to

annotate the claims made in the narratives as true or false. Their dataset has a total 35,090 words, 110 true propositions and 74 false claims. In the same domain, but for Italian language, Fornaciari & Poesio (2012) created the DECOUR - DEception in COURt - corpus. DECOUR is a corpus consisting of 3,015 utterances extracted from the transcripts of 35 hearings held in four Italian Courts. Fuller et al. (2009) collected a set of real-world data by accessing sets of data, officially known as Form 1168, provided by law enforcement personnel at military bases. At the end of the data collection process, a total of 366 written statements, 79 deceptive and 287 truthful, were obtained. Chen & Chen (2014) described to Mobile01 corpus. It is a Chinese dataset of threads, profiles and posts from the Samsung board. The spam dataset of 632,234 messages, comes from two confidential spreadsheets that appear to be internally-kept records of the spam posts.

In the last decade, there has been a growing interest in opinion spam detection. Several corpora have been developed. Lim et al. (2010) studied deceptive product reviews sourced from Amazon.com. They labeled 24 reviewers as "spammers" and 26 as "non-spammers". Li et al. (2011) studied deceptive product reviews found on Epinions.com. They labeled 6,000 reviews as either spam or not spam.

More recently, the emergence of crowdsourcing platforms as a novel and collaborative approach, has motivated NLP researchers to obtain linguistically annotated manual corpora more cost effectively. Several studies have used the Amazon Mechanical Turk (AMT) platform to this end. Rubin & Lukoianova (2015) collected a dataset of 54 stories, selfranked as truthful or deceptive, elicited using AMT. Mihalcea & Strapparava (2009) created a corpus, consisting of three datasets of true and lying texts. They collected 100 true and 100 false statements for each of the three topics (opinions on abortion, opinions on death penalty and feelings about best friends), with an average of 85 words per statement. For this purpose, the crowdsourcing platform Amazon Mechanical Turk was used to collect opinions. They also performed a manual verification of the quality of the contributions. Ott et al. (2011) have developed the first large-scale dataset containing gold-standard deceptive opinion spam. The corpus consists of 800 opinions about hotels, 400 deceptive opinions were generated by humans through the AMT platform. The other 400 truthful reviews were selected between the opinions published on TripAdvisor. Later, Li et al. (2014) expanded this corpus with positive deceptive restaurant reviews, positive reviews of hotels and positive reviews from doctors. These reviews were written by employees and experts

in each domain. The final dataset consists of 2,924 opinions.

Pérez-Rosas & Mihalcea (2014) carried out a comparative experiment to evaluate the accuracy of deception classifiers built from different cultures (United States, India and Mexico). To do that they collected three deception datasets, two in English and one in Spanish, each covering three different topics (opinion about abortion, opinion about death penalty and feelings about best friends). With this corpus of 750 statements, they showed that cross-cultural information can be beneficial to the task with accuracy ranging between 60-70%. The statements of the corpus where obtained through the AMT platform for English. For Spanish, the data were obtained through a separate web interface created specifically for this purpose. Furthermore, the experiments carried out in this work indicated that the model generated was not sensitive to different issues. For this reason, the authors conducted new experiments on open domain corpora (Pérez-Rosas & Mihalcea, 2015). In this case, they collected 7,168 sentences, 3,584 truths and 3,584 lies, through the AMT platform. These experiments confirmed that the generated model behaves adequately in the absence of a predetermined domain.

Rubin et al. (2015) manually annotated several transcriptions of the US National Public Radio. The resulting corpus consists of 144 randomly selected news.

Regardless of the concrete methodology used, the manual annotation of datasets creates gold-standard corpora. However, it has several limitations. One of the key disadvantages is that they are very expensive to obtain, but more importantly, human ability to detect deception is very poor (M. DePaulo et al., 1996). It is worth reiterating that according to Ott et al. (2011) human agreement and deception detection performance is worse than chance.

Given the limitations of manually annotated corpora, other methodologies have been studied, namely *automatically annotated corpora*. Although they do not produce a true gold-standard corpus, for some domains, they may offer an acceptable approximation. So, Jindal & Liu (2008) studied the characteristics of untruthful (deceptive) Amazon.com reviews. They implemented an approach for heuristically assigning approximate labels of deceptiveness. All duplicated reviews (different userids on the same product, same userid on different products and different userids on different products) were annotated as untruthful. Duplicate reviews were detected using the Jaccard distance similarity score. They obtained a corpus of 5.8 million reviews. Wu et al. (2010) also studied deceptive online reviews from a subset of 843 hotels from a Irish TripAdvisor dataset that comprises 29,799 reviews from 21,851

unique reviewers, covering hotels from all regions of Ireland over a two-year time window from September 2007 to September 2009. They first selected the top 41 hotels and added 5 unsuspicious ones. Users were presented with a random selection of 6 of these hotels, and were asked to mark any review that might appear suspicious. Based on the judgments provided by 55 users who completed the task, they calculated a suspiciousness score for each of the hotels.

Table 4 shows a summary of the resources cited in this section.

Work	Source	Size	Lang.	Method
Almela et al. (2012)	Opinions	600 statements	SP	M (SCE)
Chen & Chen (2014)	Web forum	632,234 messages	CH	M (SCE)
DePaulo et al. (2003)	Messages	120 samples	EN	M (UCE)
Fitzpatrick & Bachenko (2012)	-	35,090 words	EN	M
Fornaciari & Poesio (2012)	Hearing words	3,015 utterances	IT	M
Fuller et al. (2009)	Written text	366 statements	EN	M
Jindal & Liu (2008)	Amazon rev.	5.8M reviews	EN	A (SCE)
Li et al. (2011)	Epinions reviews	6,000 rev.	EN	M
Li et al. (2014)	Opinions	2,924 opinions	EN	M
Lim et al. (2010)	Amazon rev.	50 reviews	EN	M
Litvinova et al. (2016)	Written texts	1,800 documents	RU	M (SCE)
Mihalcea & Strapparava (2009)	Video and written	200 statements	EN	M (CR)
Newman et al. (2003)	Video and written	568 samples	SP	M (SCE)
Ott et al. (2011)	TripAdvisor rev.	800 opinions	EN	M (CR)
Pérez-Rosas & Mihalcea (2014)	Opinions	750 statements	EN, SP	M (CR)
Pérez-Rosas & Mihalcea (2015)	-	7,168 sentences	EN	M (CR)
Rubin & Lukoianova (2015)	Texts	54 stories	EN	M
Rubin et al. (2015)	Transc. radio	144 news	EN	M
Wu et al. (2010)	TripAdvisor rev.	29,799 reviews	En	A

Table 4: Summary of available resources for deception detection, according to the source, size, domain, language and methodology to be obtained, where M=Manual, A=Automatic, SCE=sanctioned controlled experiment, UCE=unsanctioned controlled experiment and CR=Crowfunding.

4.1.3. Systems

The first automatic proposals of deception detection were focused on the psychological or social aspects of lying. Automatic deception detection has been investigated in the context of credit card fraud (Wheeler & Aitken, 2000) and telecommunications fraud (Fawcett & Provost, 1997), among other areas. These studies share the characteristic of structured original data with predefined attributes. The natural language composition of written texts adds more complexity and ambiguity to the task of analyzing and detecting

deception in such data. Firstly, a transformation of the texts into some kind of structured format is required so that deception indicators can be captured.

At first, linguistic cues were encoded manually, and the function of the computer was limited to performing statistical analysis of those scores (Vrij, 2000) (Vrij et al., 2000) (Höfer et al., 1996) (Ruby & Brigham, 1997) (Hauch et al., 2012). Later, systems appeared that studied language and deception using a computer based text analysis program (Newman et al., 2003). In recent years, machine learning techniques for detecting deception through linguistic-based cues have been applied.

Most of the work done so far makes use of *supervised machine learning techniques*. Mihalcea & Strapparava (2009) performed deception detection via the classification problem of true and false texts. For this purpose they used Naive Bayes (NB) and Support Vector Machine (SVM), ten-fold cross validation, and a set of words as defined in LIWC (Linguistic Inquiry and Word Count[6]) (Newman et al., 2003). To tackle this task, a manually annotated dataset of 200 statements about abortion, friends and death penalty, was used. With minimal preprocessing (tokenization and stemming) and without removing stopwords, the average classification performance of 70.8% (NB) and 70.1% (SVM) was attained. The results dropped to 59.8% (NB) and 57.7% (SVM) when the portability of classifiers across topics was tested.

Fuller et al. (2009) used several different sets of linguistic-based cues as inputs for classification models. To extract the cues, a combination of GATE[7] and LIWC were used. Three classification methods were used: artificial neural networks, decision trees and logistic regression. The best results - 73.86% accuracy - were obtained with neural networks, using a single hidden layer of three nodes, and with a set of linguistic features drawn from several deception theories.

Rubin et al. (2015) analyzed the use of rhetorical structures and vector space models to detect deception. Using logistic regression and a dataset of 144 news items, they obtained an accuracy of only 56%.

Although most of the supervised systems developed have been tested for English, there have also been research work dealing with *other languages*. Fornaciari & Poesio (2012) trained models in order to perform a binary classification (false or not-false) for Italian statements issued in law courts using

[6]http://liwc.wpengine.com/ (accessed online 28 February, 2019)

[7]https://gate.ac.uk/ ((accessed online 28 February, 2019))

the DeCour corpus. They used SVM as a classifier as well as a features vector with linguistic information that made use of: the LIWC tool; the most frequent N-grams for unigrams to pentagrams of lemmas; and part-of-speech. The best result obtained was 69.37% accuracy. Almela et al. (2012) designed an automatic Spanish language classifier based on SVM for the identification of deception in written texts. The corpus used in the task consists of 600 statements, 300 true and 300 false, about three topics: opinions on homosexual adoption, opinions on bullfighting and feelings about best friends. To create the samples for the classifier, LIWC2001, the Spanish version of LIWC, was used. The system obtained an F-measure of 73.6%. Chen & Chen (2014) developed a model to detect Chinese spam opinions using the Mobile01 corpus. They obtained 61.54% of F-measure using SVM with radial basis function (RBF) Kernel and using the first post of each thread in the corpus. As features, they used bag of words, a set of features derived from basic characteristics of the contents of the posts, and information about time and thread activeness of the posts.

With the proliferation of social media and the growth of user generated content, considerable research interest in detecting true and deceptive web opinions has been generated. In this sense, several studies about *spam detection* have been carried out. Banerjee & Chua (2014) conducted a linguistic analysis using a dataset of 800 hotel reviews (Ott et al., 2011). They studied which linguistic differences in terms of readability, genre and written style could predict review manipulation. The logistic regression model obtained an accuracy of 71.25%. Kim et al. (2015) extended the linguistic features to deep semantic analysis. They proposed a frame-based deep semantic analysis method, using FrameNet (Fillmore et al., 2003) for understanding rich characteristics of deceptive and genuine opinions. The classification model using the dataset of hotel reviews (Ott et al., 2011), the subset of hotel opinions from the corpus of Li et al. (2014) and SVM, obtained an accuracy of 92,4%.

Due to scarcity of examples, some *semi-supervised machine learning approaches* have been developed. Hernández et al. (Hernández et al., 2015) employed PU-learning and Naive Bayes for building a binary classifier to detect deceptive opinions reviews. Using a dataset of 800 hotel reviews (Ott et al., 2011) they obtained 78.1% of F-measure. Moreover, they analyzed the role of opinions' polarity in the detection of deception. Their results confirmed that negative deceptive opinions are more difficult to detect that positive ones even though there are common characteristics in the way people write positive and negative deception opinions.

As with other tasks involved in NLP, some *deep learning (DL) approaches* have been analyzed. Ren & Ji (2017) explored a gated recurrent neural network model (GRNN) to learn document-level representation. Using 2,600 reviews of the total corpus (Li et al., 2014) obtained an accuracy of 83.6%. This result is not maintained in cross-domain experiments.

All the aforementioned studies consider that a text is either completely false or completely true. However, as Zhou & Zenebe (2008) argue that it is not the case. Deceptive texts combine truthful and deceptive information, or merely omit relevant details. So, an alternative is needed to traditional machine learning techniques that follow the binary paradigm. These authors applied neuro-fuzzy models that are able to tell how much of a text is deceptive and how much of it is truthful. Taking into account the part of corpus used in the experiment that only comprises written texts, the accuracy obtained using ten-fold cross validations is: 85% in emails, 67.2% in instant messages, and 69.2% in instant messages of interviews. As features, a verbal set and a small number of non-verbal behaviour cues were used.

A brief summary of the systems that address this task is presented in Table 5.

Work	Approach	Resource	Score
Almela et al. (2012)	ML (SVM)	Ad hoc	0.7360 (F1)
Banerjee & Chua (2014)	ML (LR)	Otto et al.	0.7125 (Acc)
Chen & Chen (2014)	ML (SVM)	Ad hoc	0.6154 (F1)
Fornaciari & Poesio (2012)	ML (SVM)	Ad hoc	0.6937 (Acc)
Fuller et al. (2009)	ML (NN)	Ad hoc	0.7386 (Acc)
Hernández et al. (2015)	ML (PRU-learning)	Otto et al.	0.7810 (F1)
Kim et al. (2015)	ML (SVM)	Exten. Otto et al.	0.9240 (Acc)
Mihalcea & Strapparava (2009)	ML (NB)	Ad hoc	0.7080 (Acc)
Ren & Ji (2017)	DL (RNN)	Ad hoc	0.8360 (Acc)
Rubin et al. (2015)	ML (LR)	Ad hoc	0.5600 (Acc)
Zhou & Zenebe (2008)	Neuro fuzzy models	Ad hoc	0.8200 (Acc)

Table 5: Summary of the analyzed studies on deception detection by their computational approach, training and evaluation resources and the best reported score, including the metric inside brackets: F measure (F1) and accuracy (Acc). An hyphen (-) is used where information is not provided

4.1.4. Discussion

As presented in this section, deception detection involves the identification and extraction of deception cues and the construction of a deception detection model using these cues.

A wide range of linguistic cues have been proposed and analyzed. To conclude, it is possible to establish that the same set of cues across the different contexts found on the Web, are unlikely to apply. So, although there is a small set of linguistic cues that are relevant for any task in any context, such as less complex, superfluous repetitions, or briefer messages, it is necessary to study each of these tasks and contexts independently. Moreover, some non-verbal cues, related to network behavior, could complement the knowledge provided by the linguistic cues.

Regarding models, supervised and semi-supervised machine learning approaches as well as deep learning approaches have been studied. The results obtained with these models vary between 70% and 90% accuracy. These results show the contribution of semantic information and deep learning in the task. Hence, future work should focus on how to combine both aspects in order to achieve significant progress in the task. In any case, an important fact is that the results obtained indicate that machine learning approaches are more effective at detecting deception than humans. Finally, as in all NLP tasks, corpora are needed to tackle the deception detection task automatically. In this sense, a large number of corpora have been developed. Moreover, corpora for languages other than English have been created, such as Russian, Chinese, Italian or Spanish. This is because for each team of researchers, and even for each experiment with the same researchers, a different dataset has been collected. However, there is a lack in standard annotation for the task. This fact makes it difficult to compare systems. Only in the case of opinion spam detection does it seem that a consensus has been reached and the corpus assembled by Ott et al. (2011) has been used in other related studies.

4.2. Stance detection, controversy and polarization

The task of *stance detection* concerns the use of Artificial Intelligence technologies (especially automatic learning and natural language processing) for the automatic detection of stances behind a known topic. Stance detection can be applied to two different scenarios: news and social networks. The area of stance detection in news is related to detecting misleading headlines, which involves estimating the relative perspective (or stance) of two pieces of text related to a topic, claim or issue. More specifically, the task involves classifying the stance of the body text relative to the claim made in the headline into one of the following four categories: a) *agrees* - concurrence between body text and headline; b) *disagrees* - non-concurrence between

body text and headline; c) *discusses* - same topic discussed in body text and headline, but no position taken; and, d) *unrelated* - different topic discussed in body text and headline.

The task of stance detection in a social media scenario, given a social network message (such as Twitter) and a target topic, is to automatically detect the stance (i.e., for/against) of the message regarding the topic. Most systems focus on a topic and on discovering if the associated text addresses the topic, and if so, determine whether the text is in favor, against, or indifferent to the topic.

Stance detection has recently received considerable attention and has taken off because of the rise in techniques for the detection and treatment of fake news. However, the fundamental techniques on which stance detection is based has been the focus of research for many years, as they involve applying textual entailment technology that became fashionable in the first decade of the 21st century. Hence, efforts to develop these systems have focused on revisiting the techniques of textual entailment and adapting them to the new task (and also to new resources of machine learning, such as deep learning), as well as to the construction of large resources (basically corpus) for the proper training and evaluation of these technologies.

Other researchers redefine the task as *controversy detection*, which is centered on determining if the topic of a discussion generates opposing opinions among the population. In this case, not only is the detection of contrary opinions taken into account, but also the impact on public opinion, as well as monitoring its evolution overtime. Moreover, when the task seeks to discover population clusters with generally controversial opinions on a set of different topics, the task is redefined as *polarization detection*.

The next subsections are structured as follows. Section 4.2.1 contains a detailed description of the resources and competitions found in the systematic review regarding stance, polarity and controversy detection. Sections 4.2.2 and 4.2.3 reported the research studies regarding stance detection in news and social media, respectively. Section 4.2.4. presents research regarding polarization and controversy detection. Finally, section 4.2.5 shows some conclusions about the different research presented, as well as the possible future research lines in this subtask.

4.2.1. Resources and competitions

New machine learning techniques applied to the task of stance detection need to have large datasets for their proper training. Powerful systems,

such as those based on deep learning have shown very promising results for many facets of ML, but they are unlikely to be competitive with traditional ML technologies when they are trained with low or medium sized datasets (Hanselowski, 2018). For this reason, the construction of resources for the training of these systems is one of the priority tasks that needs to be addressed.

Bowman et al. (2015) made a significant contribution in this area with the development of *Stanford Natural Language Inference (SNLI) corpus*, a collection of labelled sentence pairs, written by humans based on image captioning. This corpus has 570,152 sentence pairs labelled for entailment, contradiction, and semantic independence, collected by means of the Amazon Mechanical Turk. The corpus was evaluated by testing simple lexicalized models as well as neural network models achieving near 80% success in both cases. Furthermore, in this experimentation, specific approaches for stance detection were tested, as well as approaches for Recognizing Textual Entailment (RTE) challenge tasks. It was found that both approaches are based on the same technology.

Resource building for stance detection has also benefited from the work *Emergent dataset* developed by Ferreira & Vlachos (2016). In this case, they set themselves the goal of obtaining data for classifying the stance of a news article headline with respect to its associated claim. In this way, Emergent, is a dataset containing 300 rumoured claims and 2,595 associated news articles, collected and labelled by journalists with an estimation of their veracity (true, false or unverified). This dataset has been extracted from the Emergent Project (Silverman, Visited January, 2019), a rumour debunking project.

The launching of the first challenges related to fake news analysis, and specifically, to the task of stance detection was triggered by a demand for technologies that analyze fake news and the increasing availability of annotated corpora. Next, an analysis follows of two of the most recent challenges set that have greatly impacted the scientific field.

Babakar et al.[8] presented the *Fake News Challenge FNC-1*, using Ferreira & Vlachos (2016) as a starting point. FNC-1 aims to compile a gold standard to explore Artificial Intelligence technologies, especially machine learning and natural language processing, applied to detection of fake news.

[8]http://www.fakenewschallenge.org/ (accessed online 28 February, 2019)

To carry out this macro-challenge, the organizers decided to start with stance detection. In this case, the FNC-1 dataset with 50,000 headlines (1,600 different news items) was released. These headlines were classified as follows: agree (7.4%), disagree (1.7%), discuss (17.8%), unrelated (73.1%). The competition received a total of 200 submissions achieving scores of around 82% in the best ranked submissions. The organization proposed a simple baseline using hand-coded features and a gradient boosting classifier, available at Github[9]. The baseline implementation also includes the following: code for pre-processing text; splitting data carefully to avoid bleeding of articles between training and test; k-fold cross validation; scorer; and, most of the CRUD[10] needed to carry out the experiment. The hand-crafted features include word/ngram overlap features, and indicator features for polarity and refutation. With these features and a gradient boosting classifier, the baseline achieves a weighted accuracy score of 79.53% (as per the evaluation scheme described above) with a 10-fold cross validation.

Also, the SemEval challenge addressed the problem of stance detection, although using a different application: *SemEval-2016 task 6: Detecting stance in Tweets* (Mohammad et al., 2016). This task is formulated as follows: in a given tweet text and a target entity (person, organization, movement, policy, etc.), NLP systems must determine whether the tweet is in favor or against the given target, or whether neither inference is likely. This task permitted two frameworks: i) *supervised* where a previously tagged training collection was provided, ii) *poorly supervised* where tagged training was not provided. As a result, a dataset with 4,870 English tweets for stance detection was released. The dataset addressed six targets: Atheism, Climate Change is a Real Concern, Feminist Movement, Hillary Clinton, Legalization of Abortion, and Donald Trump. The challenge received 19 submissions for the task in a supervised framework and 9 for the unsupervised framework, and the best scores were around 67% and 56% respectively, from systems that employed a wide array of features and resources.

The previously discussed challenges and resources have made it possible to train an important set of systems specifically addressing two scenarios: i) stance detection in news, and ii) stance detection in tweets. The following

[9]https://github.com/FakeNewsChallenge/fnc-1-baseline (accessed online 28 February, 2019)

[10]Create, Read, Update and Delete

subsections focus on each of these scenarios.

4.2.2. Stance detection in news

Research studies dealing with *Stance Detection in news* have followed a variety of approaches that range from classical methods of machine learning to the most advanced deep learning models.

The use of *neural networks* has always played a fundamental role in the development of artificial intelligence systems, and as expected, they also play a decisive role in the task of stance detection. Using this technology, the previously mentioned work (Bowman et al., 2015) tested the viability of the SNLI corpus by means of a neural network model centered around a *Long Short-Term Memory network* (LSTM RNN model), through which they obtained a score of 77.6%.

The evolution of these systems towards more complex neural networks has led to models such as *multilayer perceptrons* (MLPs) that improved the results. This is the case of systems such as the one presented by Hanselowski (2018) Athene in the FNC-1 challenge, which reached a score of 81.97% and became the second best team in the competition. The one presented in Riedel et al. (2017) achieved a score of 81.71% and ranked third in the same challenge.

Some of the most evolved neural networks to date are the *convolutional neural networks* (CNN), the deep learning systems defined as an extension (or rather a supertype, as according to some authors) of the MLPs. Based on this technology, the work presented by Baird et al. (2018) applied a one-dimensional convolutional neural net on the headline and body text, represented at the word level using Google News pretrained vectors. The output of this CNN is then sent to a multilayer perceptron with 4-class output: "agree", "disagree", "discuss", and "unrelated", and trained end-to-end. Using this combination CNN - MLP, the system outperformed all the submissions and achieved a score of 82.02%, and the first position in the FNC-1 challenge.

The work presented by Chaudhry et al. (2017), using the framework defined at FNC-1, develops several neural network-based models to tackle the problem, concluding that the LSTM-based bidirectional conditional encoding model using pre-trained GloVe word embedding is the best solution with a performance of 97% in classification accuracy.

However, the work of Rakholia & Bhargava (2016) focuses discussion on a new area. After applying several existing neural network architectures to the problem of stance detection in news articles over the FNC-1 dataset, they

showed that LSTM-based neural network architectures (84.96%) or simple feedforward neural nets (85.61%) perform better than CNN (68.28%) due to the difficulty of fine-tuning in deep learning processes.

For the reasons described above, there are many authors who still opt for classical methods of *machine learning*. Indeed, we can assert that lexicalized models perform these tasks and achieve values comparable to other more evolved systems. This is one of the conclusions reached by Bowman et al. (2015) when they tested a lexicalized classifier to evaluate the SNLI corpus and reached a score even higher (78.2 %) than those obtained with the more evolved LSTM RNN model (77.6 %). Additionally, the corpus presented in Ferreira & Vlachos (2016) was evaluated by a 3-way classification task using a logistic regression classifier (Maximum Entropy model). This evaluation took into account features extracted by combining the headline and the claim, obtaining scores of 73%.

Regarding other languages, the work presented by Wei & Wan (2017) divided the headlines into ambiguous and misleading, and treated them separately. They work with Chinese, and in order to identify ambiguous headlines they used Support Vector Machines, a set of basic features and a set of linguistic patterns. For misleading headlines, they considered independent and dependent body features, such as informality, sentiment, similarity and textual entailment. They used a co-training approach for the classification. A corpus of 40,000 pieces of news from Chinese news sites was created, with scores of 80% accurate headlines and of 20% misleading and ambiguous headlines. They obtained an F-measure of 72.4% using the co-training approach.

The work of Bourgonje et al. (2017) that was evaluated on the FNC corpus showed that the use of simple ML techniques returned very satisfactory results (89%) that greatly improved the scores of other much more complex systems presented to the FNC-1 challenge. In this case, a simple lemmatisation-based n-gram matching was used for the binary classification of *related* vs. *unrelated* headline/article pairs by means of logistic regression.

Studies such as Rubin et al. (2016) are focused on providing a conceptual overview of satire and humour, discovering and highlighting the unique features of satirical news that tend to be adopted in both the format and style of this type of journalistic reporting. They used a feature combination (Absurdity, Grammar and Punctuation) to detect satirical news with a 90% precision and 84% recall (F-score=87%). However, satirical news is not considered as fake news, and therefore is beyond the scope of this section.

To conclude, the methods based on *deep learning* obtain promising results in the experiments carried out. However, they need fine-tuning and, in many cases, they need to be complemented with traditional ML resources to achieve the results of classical techniques.

A brief summary of the systems that address this task is presented in Table 6.

Work	Approach	Resource	Score
Dias & Becker (2016)	Hybrid (CNN/DT)	FNC1	0.8202 (FNC1)
Bourgonje et al. (2017)	ML (LR)	FNC1	0.8959 (FNC1)
Bowman et al. (2015)	DL (RNN-LSTM)	SNLI	0.776 (Acc)
Chaudhry et al. (2017)	DL (LSTM-GRU cells)	FNC1	0.97 (FNC1)
Ferreira & Vlachos (2016)	ML (Max. Entr.)	Emergent	0.73 (Acc)
Hanselowski (2018)	DL (separate MLPs)	FNC1	0.8197 (FNC1)
Rakholia & Bhargava (2016)	DL (LSTM-Feed Forw.NN)	FNC1	0.8561 (FNC1)
Riedel et al. (2017)	DL (MLP)	FNC1	0.8172 (FNC1)
Wei & Wan (2017)	ML (SVM)	Ad hoc	0.7240 (F1)

Table 6: Summary of the analyzed studies on stance detection in News according to the computational approach, the resource used to train and evaluate and the best score reported, including the metric inside brackets: Fake News Challenge-1 score (FNC1), F measure (F1) and accuracy (Acc). An hyphen (-) is used where information is not provided

4.2.3. Stance detection in tweets

Stance detection in tweets has similarly been approached from different perspectives. However, unlike stance detection in news, trainers cannot count on the pair headline and claim but only headlines (tweets). The SemEval2016 competition (task 6) became one of the best launch pads for the take-off of these systems.

Neural networks are among the most used techniques in stance detection in news, playing a decisive role in this task. This is the technique used by Zarrella & Marsh (2016) who applied a *recurrent neural network* organized into four layers of weights to tackle the SemEval2016 task 6A (supervised identification of stance in tweets). The RNN was initialized with features learned via distant supervision on two large unlabelled datasets. In parallel, they trained embeddings of words and phrases with the word2vec skip-gram method. Finally, they used the features to learn sentence representations via a hashtag prediction task. These experiments ranked first among the 19 systems with F-score of 67.8%, reporting 71.1% in a non-official test.

Wei et al. (2016) addressed the supervised task of SemEval2016 through a specific *convolutional neural network* for stance detection. In this case,

the input was the learning of the word embedding model extracted from Google News database. Afterwards, the CNN model was trained with the SemEval2016 Task 6 dataset. Finally, they used a vote scheme to predict the label of the test set. This submission ranked second in the task A (supervised) with F-score of 67.33%, and first in the task B (unsupervised) with a strategy to build the necessary training dataset since it was a supervised method (F-score = 56.28%).

Another approach is the one that tackles the task from the point of view of a *combination of basic algorithms optimized with genetic algorithms*. This is the case with the work of Tutek et al. (2016), who used an ensemble of learning algorithms that were fine-tuned by using a genetic algorithm. These experiments revised a long set of classifiers: Support Vector Machine (SVM); Random Forest (RF); Logistic Regression (LR); Gradient Boosting (GB); Multinomial Bayes (MB); Extra Trees (ET); and, general stochastic gradient descent classifier (SGDC). To build the model, standard lexical and task-specific features were employed. The system also participated in SemEval2016 task 6A, ranking third position with F-score 66.83%.

The traditional classification methods supervised by means of resources have also been used to address task B of SemEval (unsupervised or poorly supervised framework). This is the case of the work by Dias & Becker (2016) who automatically generated a training corpus by means of a rule-based system. Then a *SVM classification algorithm* was trained with it. They ranked third in the SemEval2016 task 6B with F-score 42.32%.

Also the work of Krejzl & Steinberger (2016) is based on this type of technique. In this case, a *maximum entropy classifier* that was trained with mainly surface-level, sentiment and domain specific features extracted from entity-centered sentiment dictionaries and domain stance dictionaries. They ranked fourth in SemEval2016 task 6B with F-score 42.02%.

Table 7 presents a summary of work that has addressed the task of stance detection in tweets.

4.2.4. Polarization and controversy detection

Some systems have extended the stance detection task by focusing on perceiving the political polarization between users, generally within a social network. This is a problem that can be addressed by adapting the specific models for stance detection, or approaching the problem from totally different perspectives, where the position of each user is not analyzed for a certain issue, but groups of users with similar positions are detected. Thus, in Finn

Work	Approach	Resource	Score
Dias & Becker (2016)	ML (SVM)	SemEval2016	0.4232 (F1) (unsupervised)
Krejzl & Steinberger (2016)	ML (MaxEntr)	SemEval2016	0.4202 (F1) (unsupervised)
Wei et al. (2016)	DL (CNN)	SemEval2016	0.6733 (F1) (supervised) 0.5628 (F1) (unsupervised)
Zarrella & Marsh (2016)	DL (RNN)	SemEval2016	0.6780 (F1) (supervised)

Table 7: Summary of the analyzed studies on stance detection in Tweets by their computational approach, training and evaluation resources and the best reported score. , including the metric inside brackets: F measure (F1). An hyphen (-) is used where information is not provided

et al. (2014) a proposal that analyzes the co-retweeted network among Twitter users based on *graph models* is presented, using the Gephi ForceAtlas2 for implementation. The tweets are collected by searching for keywords related to a certain event. Specifically, the US presidential debate between Obama and Romney in October 2012 was used. These same keywords were used to detect supporters of both contenders. The researchers used the Co-Retweeted Network that was constructed as the undirected weighted graph connecting accounts which have been retweeted by members of the audience. Finally, the computation of the polarity of the event itself as well as the polarity of the major accounts participating in the discussion were shown.

Although the above cited studies could be evaluated extrinsically by the users of the graphical representation system, none of the previous studies have proposed an intrinsic and comparable evaluation method for the task. Thus, in order to address the issue of providing a comparable metric, some studies have simplified the complex task of identifying controversy by reducing it to a simple binary classification of polarity with respect to a controversial topic. One of the most relevant work in this sense is the one carried out by Dori-Hacohen & Allan (2015) who used a *simple nearest neighbors classifier* (kNN) to polarize the controversy in Wikipedia articles and conducted a comparative study of the results against different baselines. In this case the authors reported a F-score of 65% (accuracy=73%) with a gain increase of 20% over their baselines. This work was later improved by Jang et al. (2016) who implemented a version of the previous kNN algorithm based on Probabilistic Language Modelling to confront the Wikipedia controversy with results of 83.5% accuracy.

In the same way, to address the evolving problem of polarization, it is possible to detect when the topic of a discussion is controversial by monitor-

ing the topic's behavior. In this scenario, the work presented by Garimella et al. (2018) is also based on graph models. This paper proposes three phases: (i) building a conversation graph about a topic taking into account relationships such as: retweets, followers, and content similarity; (ii) partitioning the conversation graph to identify potential sides of the controversy; and (iii) measuring the amount of controversy from the characteristics of the graph.

Table 8 presents a summary of studies that have addressed the task of controversy/polarization detection.

Work	Approach	Resource	Score
Dori-Hacohen & Allan (2015)	ML (KNN)	Wikipedia	0.73 (Acc)
Finn et al. (2014)	Graph Models	Co-Retweeted Net	-
Garimella et al. (2018)	Graph Models	Twitter	-
Jang et al. (2016)	ML (Prob. LM)	Wikipedia	0.8350 (Acc)

Table 8: Summary of the analyzed studies on controversy detection by their computational approach, training and evaluation resources and the best reported score, including the metric inside brackets: Accuracy (Acc). An hyphen (-) is used where information is not provided

4.2.5. Discussion

The following conclusions can be drawn from the results analysed in this section in relation to stance detection and controversy detection. First, with respect to *stance detection in news*, the current systems are based on detecting the entailment between the headline and the news, and they are working with relatively high scores (around 85%-95%). Among these, the DL systems are obtaining the best results, although well adjusted ML systems also provide very competitive results. However, (Hanselowski et al., 2018) demonstrated that stance detection is a challenging problem because they evaluated the performance of the three top-scoring systems at FNC-1 with a new dataset called Argument Reasoning Comprehension (ARC) (Habernal et al., 2018), and the systems were not able to resolve difficult cases, detecting a gap in the research were argumentation is involved.

Second, the task of *stance detection in tweets* becomes a more complex task as it lacks the headline/body pair to train and make decisions. In this sense, when the task is performed with supervised training, all systems achieve similar results (around 65%-70%), but clearly providing an opportunity for future research to improve the score. Moreover, few differences

were demonstrated between DL and ML approaches. However, in the case of unsupervised work, the results are exceptionally low (40%-55%) with many aspects remaining unresolved.

Third, the problem associated with the *detection of controversy* is best understood as a multidimensional version of the stance detection problem, in which the same topic can generate multiple disagreements in different directions forming opinion clusters. So far this problem has been approached from two different points of view. On the one hand, the problem of multi-dimensionality has been approached through the graphical representation of the problem, without measuring the degree of disagreement between the different positions. On the other hand, an attempt has been made to measure the degree of disagreement by reducing the problem from controversy to a *polarization problem*, that is, turning it into a mono-dimensional problem that returns us to the problem of stance detection. However, the calculation of polarity is a very simplified version of the original problem of controversy, and therefore there is much room for research in a task that is still in its infancy.

To conclude, Table 9 shows a summary of the resources cited in this section.

Work	Source	Size	Subtask
EMERGENT(Ferreira & Vlachos, 2016)	News	2,595 news	stance
Finn et al. (2014)	Twitter	1,895,334 tweets	controversy
FNC1:2016 (Babakar et al., 2016)	News	49,972 headline-body	stance
SemEVAL(Mohammad et al., 2016)	Twitter	4,870 tweets	stance
SNLI(Bowman et al., 2015)	Image Captions	570,152 sentence-pairs	stance
Wei & Wan (2017)	News	40,000 news	stance

Table 9: Summary of available resources for stance detection, controversy and polarity, according to the source, size and the concrete subtask addressed

4.3. Automated Fact Checking

Computational or automated fact checking consists of applying existing artificial intelligence technology to automatically check the veracity of a public claim against all the available data, and classify it according to a veracity value (Dale, 2017). These values are known as a 5-point veracity scale (Vlachos & Riedel, 2014) and they are explained next:

- **True:** The statement is accurate and there is nothing significant missing.

- **Mostly True:** The statement is accurate but needs clarification or additional information.
- **Half True:** The statement is partially accurate but leaves out important details or takes things out of context.
- **Mostly False:** The statement contains an element of truth but ignores critical facts that would give a different impression.
- **False:** The statement is not accurate.

Some authors include an additional point in the veracity scale known as "Pants on-fire False",[11] meaning the statement is not accurate and makes a ridiculous claim.

The main objective of this task is to enhance the fact checking role and make it more accurate. This releases the journalist's time for more interpretative reporting (providing context, analysis, and possible consequences of events). In all the aim of fact checking is to limit the spread of unsubstantiated claims. According to FullFact organization (the UK's independent fact checking charity), the automated fact checking process is divided into the same stages as the manual fact checking performed by journalists or researchers (FullFact.org, 2016).

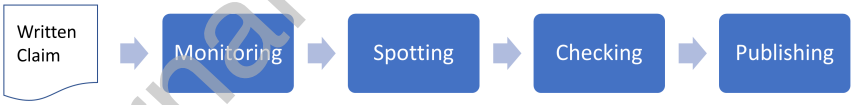


Figure 2: Four stages required to perform Fact checking procedure

As presented in Figure 2 the fact checking process consists of four subtasks. The first subtask involves *monitoring* different media and social networks, such as Twitter, to extract the possible claims. After this, in the *spotting* subtask the relevant claims are determined. Then, the *checking* subtask classifies the claim according to its veracity, and the last subtask consists of *publishing*, presenting the previously extracted content in a human-readable and human friendly format.

[11] https://www.politifact.com/truth-o-meter/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/ (accessed online 28 February, 2019)

An analysis follows of the main challenges involved in each of the four subtasks (FullFact.org, 2016):

Monitoring: Determines the whole context of the claim, which includes automatically identifying the sender/speaker, the receiver about who or what the message concerns, as well as the temporal and spatial framework of the claim.

Spotting: In this subtask, it is important to: i) identify previous fact-checked claims in new text, ii) identify new factual claims that have not been fact-checked before, iii) make judgments about the priority of the claims, and iv) deal with paraphrasing.

Checking: There are three working approaches to automated checking:

Reference approaches: use built-in knowledge (known sources and existing fact-checked claims) to determine the veracity of a claim.

Knowledge graph approaches: use canonical information extracted from graphs to check the veracity of the claim.

Contextual approaches: use social and other contexts related to the claims, which involves considering issues, such as: the reaction claims get on the social web; or how they spread; and, the controversy that they generate. With these approaches, it is not possible to determine if a claim is true, but they are necessary to discover the likelihood of the claim's veracity.

Creating and publishing: This subtask is related to automated journalism and it is the reverse task of fact checking. It involves applying Automatic Language Generation techniques to the structured data previously obtained in the previous three subtasks. Common standards are already emerging for presenting fact checks through work done by schema.org[12] based on the work that human fact checkers and journalist already do.

The following section 4.3.1 presents the different datasets and resources found in the literature regarding automated fact checking. Then, section 4.3.2 involves briefly presenting systems dealing with monitoring and spotting as well as a more in-depth focus on the checking systems. Thus, within the checking sub-group, we have divided the studies into reference approaches, knowledge graph approaches and contextual approaches. Finally, in section 4.3.3 a brief summary of the subtask and the state-of-the-art is presented, as well as a contrast between approaches and the gaps to be tackled by future

[12]https://github.com/schemaorg/schemaorg/issues/1061 (accessed online 28 February, 2019)

research.

4.3.1. Resources and competitions

The automated fact checking task lacks standard annotated datasets. Therefore, different authors have proposed their own datasets. Multiple datasets, which can be sourced from manual fact checkers such as Politifact[13] and Full Fact,[14] have been created by authors for machine learning or evaluation purposes (Vlachos & Riedel (2014), Rashkin et al. (2017), LIAR (Wang, 2017)). These corpus are annotated with the veracity values used in manual fact checking (see section 4.3).

Emergent corpus (Ferreira & Vlachos, 2016) is a dataset of 300 rumoured claims, and 2,595 associated news articles, labelled by journalists with a veracity tag (true, false or unverified), providing a real-world data source for human language technology tasks in the context of fact-checking. This corpus could also be applied for stance detection tasks, as subsequently explained.

In addition, task driven competitions provide a source of datasets and establish the evaluation framework of the task. One of the competitions in this field was the Fast and Furious Fact Check Challenge (HeroX fact checking Challenge)(Francis & Fact, 2017) completed in January 2017. The annotation of the claims was done with a four-grade scale (true/false/somewhat true/somewhat false). The competition set a minimum threshold of 80% accuracy, which was not achieved by any of the participating teams. The Sheffield team (Thorne & Vlachos, 2017), Claimbuster(Hassan et al., 2017a) and the Ovidiu Dobre team were the three best teams in this competition.

There was also a Hackathon called FactHack, [15] held in 2017 with the aim of developing tools to help scale, target, and evaluate the fact checking task. The hackaton objectives were: i) to detect already fact-checked claims appearing in different contexts, and ii) to improve their live factchecking systems by building a system that immediately finds matches with claims already checked (and the accompanying verdicts). For the hackathon they decided to break the problem down into 3 key areas that they needed help with: i) Real-time search, ii) Pre-processing of numbers, and iii) Stacked tokens.Firstly, the real-time search objective was accomplished by means of

[13]http://www.politifact.com (accessed online 28 February, 2019)

[14]http://www.fullfact.org (accessed online 28 February, 2019)

[15]https://fullfact.org/blog/2017/jan/facthack-our-hackathon-facebook-flax/ (accessed online 19 July, 2019)

Luwak query engine, as it enables the search of a rapid stream of documents to find factual claims. Secondly, regarding preprocessing, the team had made some important decisions, identified some awkward edge cases, and started to think about how they could adapt the system to other countries. Finally, stacked tokens, as the most experimental aspect of the project, implies being able to automatically detect phrases like something is rising and determine where something is a noun phrase like "crime" for example. The team of Solr experts in the Hackathon worked on this and provided the organizers with the ability to work out how to take the next steps towards ever more nuanced types of searches.

In 2018, the two main competitions were: i) the CLEF-2018 Fact checking Lab[16] for the automatic identification and verification of claims in political debates(Nakov et al., 2018). The dataset delivered by this competition was obtained from `FactCheck.org` and it annotates statements with true/half-true/false values (Barrón-Cedeño et al., 2018); and, ii) the Fact Extraction and VERification (FEVER) [17], which is a workshop on fact extraction and verification providing a dataset of 220K claims verified against Wikipedia (Thorne et al., 2018b) (Thorne et al., 2018a). The statements in the corpus were annotated with supported/ refuted/ notenoughinfo. The best performance in this challenge was obtained by UNC-NLP team(Nie et al., 2018) with a FEVER score value of 64.21 %.

Regarding datasets and competitions with veracity value annotation, RumourEval (Derczynski et al., 2017) is a SemEval shared task that aims to identify and handle rumours and reactions to them, in text. They present an annotation scheme and a large dataset (Pheme) covering multiple topics. They have a task related to predicting veracity of rumours (Subtask B - Veracity prediction) with 330 annotated Tweets with true/false values. The dataset is also covering stance classification but this beyond the scope of this section.

Table 10 presents a summary of the datasets addressed in this section for automated fact checking. Some of the datasets addressed in this table are built by the authors, and some of them are derived from related competitions. The number of different values for labels in the annotation of the resources is also indicated in the table.

[16]http://alt.qcri.org/clef2018-factcheck/ (accessed online 28 February, 2019)

[17]http://fever.ai/ (accessed online 28 February, 2019)

Work	Source	Size	Label
CT-FCC(Barrón-Cedeño et al., 2018)	FactCheck.org/Snopes	300 claims	three-grade
Emergent(Ferreira & Vlachos, 2016)	News articles	300 claims	three-grade
FEVER(Thorne et al., 2018a)	Wikipedia	185K claims	three-grade
HeroX(Francis & Fact, 2017)	Full Fact	90 claims	four-grade
LIAR (Wang, 2017)	PolitiFact	12.8K claims	six-grade
Pheme(Derczynski et al., 2017)	Tweets	330 claims	true/false
Rashkin et al. (2017)	PolitiFact	4,366 claims	six-grade
Vlachos & Riedel (2014)	Channnel4/PolitiFact	221 claims	five-grade

Table 10: Summary of available resources for automated fact checking by source, size and type of data labelling (annotation)

4.3.2. Automated Fact Checking systems

In this section a map of the main studies regarding automated fact checking are presented and compared. We focus on monitoring, spotting and checking since these are the subtasks that are presently being tackled automatically, even though the state of the art is highly dynamic.

The main systems focused on *monitoring and spotting* are:

Emergent.info [18] is a real-time rumour tracker, the result of a research project from the Tow Center for Digital Journalism at Columbia University and it focuses on how unverified information and rumour are reported in the media. For a given a claim, the website shows a veracity assessment –'true', 'false' or 'unverified'–. Additionally, the claim's source and social dissemination, as well as the media outlets that published the claim are provided. Emergent needs a lot of manual input and, furthermore, the system automatically detects misleading headlines and the evolution of rumours over time.

The Contentcheck [19] (Cazalens et al., 2018) (Manolescu, 2017) is an ongoing project that consists of a claim detection and an analysis tool designed by Le Monde newspaper and a French academic consortium. They characterized automated fact checking as a content management problem, drawing from data and knowledge management, natural language processing and information retrieval. Machine learning is also leveraged for many of the tasks involved. They also considered important additional information on to how or why a claim is disputed, such as related articles or argumentation graphs.

[18] http://www.emergent.info/ (accessed online 28 February, 2019)

[19] https://team.inria.fr/cedar/contentcheck/ (accessed online 28 February, 2019)

FactWatcher (Hassan et al., 2014) is a system that helps journalists to monitor facts which may serve as leads to news stories. Given an append-only database, on the arrival of a new tuple, the system monitors if the tuple triggers any new facts. FactWatcher also includes fact ranking, fact-to-statement translation and keyword-based fact search.

Wu et al. (2014) proposes a framework that models claims based on structured data as parameterized queries. This framework allows the formulation of practical fact-checking tasks as computational problems.

Research focused on *checking* techniques are classified according to the evidence used by the systems for determining the veracity of a claim. According to Thorne & Vlachos (2018), most systems are performing supervised machine learning approaches but using different evidence in the process. The classification of the systems based on the evidence used is presented next.

Reference approaches. Systems based on reference approaches use reasoning with known sources, and previously fact-checked claims.

This reduces the task to one based on textual similarity, as is the case with Full Fact (FullFact.org, 2016), which has been operating on building scalable, robust, automated fact checking tools to be used in newsrooms and by fact checkers globally since 2013.[20] The platform aims to cover all the subtasks in fact checking and uses industry standard media monitoring software for claim recognition. Furthermore, pattern recognition and structured data is used for statistical claims identification and statistical fact checking. In the future, 'Robocheck' will be a platform that combines human and automated fact checking.

In Vlachos & Riedel (2014) also defined the fact checking task as determining the semantic similarity between statements, which were previously annotated by fact checking agencies and journalists so as to classify the veracity of stories. In the case of Rashkin et al. (2017), evidence beyond the claim is not considered. They use these manually fact-checked claims to extract linguistic features from LIWC, and then apply ML approaches to determine the veracity of the text.

Karadzhov et al. (2017) present a framework that uses external sources, tapping the potential of the entire Web as a knowledge source. The combination of the representational power of neural networks with the classification

[20]https://fullfact.org/automated (accessed online 28 February, 2019)

of kernel-based methods results in a very strong performance. They used part of the rumor detection dataset created by Ma et al. (2016).

Claimbuster (Hassan et al., 2017a, 2015, 2017b) is a fact checking platform that performs factual claim spotting in political discourses. To perform the task, it uses natural language processing and supervised learning over a human-labelled dataset of check-worthy factual claims from the U.S. general election debate transcripts. They classify sentences into three types: non-factual sentence (NFS), unimportant factual sentence (UFS) and check-worthy factual sentence (CFS). A set of features are used and applied to different supervised learning methods, including Multinomial Naive Bayes Classifier, Support Vector Machine, and Random Forest Classifier. The best performance was obtained with Support Vector Machine.

The main problem with systems based on reference approaches is that they do not consider evidence beyond the text of the claim, and a sentence that appears to be credible may be inherently false. Furthermore, to be checked effectively, the claim would have to be manually fact-checked or to be contained in a source.

Knowledge graph approaches. Knowledge graphs are a structured and canonical format that provide rich world knowledge, and support the fact checking task. It is important to underscore that a common input to fact checking approaches that facilitates access to (semi-)structured knowledge is subject-predicate-object triples.

Ciampaglia et al. (2015, 2018); Shiralkar et al. (2017); Shao et al. (2016) present an unsupervised network-flow based approach to determine the truthfulness of a statement of fact, expressed in the form of a (subject, predicate, object) triple. They took a knowledge graph (KG) of real facts to match the claims and return a truth score, indicating the probability of the claim's accuracy. To perform this, they use two methods called Knowledge Stream and Relational Knowledge Linker. *Knowledge Stream* is based on network flow and employs multiple short paths. *Relational Knowledge Linker* finds the single shortest path. Moreover, they proposed a method to measure the similarity between any two relations purely based on their co-occurrence in the KG. Many KGs (e.g., YAGO2 (Hoffart et al., 2013) and Wikidata (Erxleben et al., 2014)) now contain facts enhanced by spatio-temporal details. Checking the veracity of facts during a restricted time frame or at a specific location is another important challenge.

In the work presented by (Vlachos & Riedel, 2015) (Thorne & Vlachos,

2017), a framework to verify numerical claims that uses a distantly supervised machine learning approach is presented. They identify surface patterns in text, which describe relations between two entities in a knowledge graph. At first, they focused on performing fact checking of simple claims about statistical properties, such us *"population of Germany in 2015 was 80 million"*. These types of numerical claims comprised a quarter of the test instances in the Fast and Furious Fact Check Challenge (HeroX Challenge). In recent research they have extended the system to use temporal expressions and knowledge bases consisting of multiple tables to improve the fact checking task. Their fact checking process has three steps. Firstly, named entities in the claim are linked to entities in their Knowledge Base (KB) and a set of tuples that includes the entities found are retrieved. Secondly, these entries are filtered in a relation matching step. Using the text in the claim and the predicate as features, they classify whether the tuple is relevant. And finally, the values in the matched tuples from the KB are compared to the value in the claim. A verdict is deduced if a certain threshold is surpassed. The system was tested on a published data set (Vlachos & Riedel, 2014) (Vlachos & Riedel, 2015), in the HeroX fact checking challenge[21] and FEVER (Thorne et al., 2018b).

Contextual approaches. Finally, within the checking subtask there is considerable research adopting contextual approaches. This means reasoning about the social and other contexts related to the claims, which involves considering issues, such as the reaction claims get on the social web, or how they spread, the controversy that they generate, and so on.

In this sense, Wang (2017) incorporates metadata such as the originator of the claim, speaker profile and the media source of the claim. This approach also matches the claims with already fact-checked claims. Following the same idea, Long et al. (2017) consider a more extended profile of the originators of the claims to deduce the verdict. It also includes a credibility measure that takes into account how often the originator's claims are false.

Pheme project (Derczynski & Bontcheva, 2014) is a research project using contextual information to identify four kinds of false claims in social media and the web in real time. Claims are classified as rumours, disinformation, misinformation or speculation. The Pheme project considers temporal and spatial contexts, in addition to information about who sent or shared the

[21] https://www.herox.com/factcheck (accessed online 28 February, 2019)

claim, the user stance, and so on. Another work using this approach is the TwitterTrails project (Metaxas et al., 2015), which enables members of the media to track the trustworthiness of stories shared on Twitter.

A summary of the systems included in the automated fact checking section are presented in Table 11. Evaluation results of the systems are provided but they are not directly comparable since the resources used for their evaluation are different. Some evaluation results were not reported in literature.

Work	Approach	Resource	Score
Ciampaglia et al. (2015)	KG+ML(RFC,k-NN)	Ad hoc	0.990 (F1)
Hassan et al. (2017a)	Ref.+ML(NBC,SVM,RFC)	Human-labeled	0.818 (F1)
FullFact.org (2016)	References+ML	-	-
Karadzhov et al. (2017)	Ref.+ML(NN,SVM)	Ma et al.	0.772 (F1)
Long et al. (2017)	Context+ML(LSTM+Att)	LIAR	0.415 (Acc)
Derczynski & Bontcheva (2014)	Contextual	Social Networks	-
Rashkin et al. (2017)	LIWC+ML(LSTM)	Rashkin	0.560 (F1)
Shiralkar et al. (2017)	Knowledge Stream	Ad hoc	0.9163 (AUROC)
Thorne & Vlachos (2017)	KG+ML	HeroX	0.6818 (Acc)
Thorne et al. (2018a)	Decomposable Attention	FEVER	0.3187 (Acc)
Metaxas et al. (2015)	Contextual	Twitter	-
Nie et al. (2018)	Neural Semantic Network	FEVER	0.6421 (Acc)
Wang (2017)	Context+ML(SVM,CNN)	LIAR	0.270 (Acc)

Table 11: Summary of the analyzed studies on automated fact checking according to the computational approach, training and evaluation resource and the best score reported, including the metric inside brackets: area under ROC curve (AUROC), F measure (F1) and accuracy (Acc). An hyphen (-) is used where information is not provided

4.3.3. Discussion

As presented in this section, fact checking involves four stages, beginning with monitoring claims for checking and concluding with a final report that justifies the veracity value of a claim. Although there are some attempts to create platforms that perform the entire process, such as VERA (Ba et al., 2016) or Claimbuster (Hassan et al., 2017a), this fully automated fact checking task is still far from being accomplished. First it is necessary to determine the claims to be checked and what information can be retrieved from them, or even parameterized. The systems that are focusing on the checking stage are systems that mostly use machine learning techniques to classify a claim into a veracity value, although they are based on different evidences (references, knowledge graphs or contextual information).

As with other tasks involved in fake news detection, the comparison between systems is unreliable, because although the results reported are mostly

measured with F-measure, AUROC or Accuracy, many studies use author created datasets to evaluate their systems making comparison between them difficult. Depending on the evidence, some conclusions can be drawn from the summary table. In the case of reference approaches, most report the F-measure score, but using different evaluation datasets, reaching ceilings of between 77% - 82%. In the case of systems based on knowledge graphs, when datasets are created by the authors themselves they seem to produce higher results. However, accuracy results decrease when datasets are created by challenges such as HeroX and FEVER. In the Herox challenge, a minimum threshold of 80% accuracy was established, and none of the participating systems managed to reach that threshold. In the case of FEVER, Thorne et al. (2018a), the organizers, gave a baseline value (about 32%) that was doubled by the system with best results (about 65%), but still being relatively low values that imply an improvement niche for future research. These accuracy values are even lower if we talk about contextual approaches, where systems evaluated with the LIAR corpus are below 50% accuracy. These results reveal how difficult the task is and how far it is from being fully resolved. Apart from this, the results suggest that the use of a combination of evidence sources, as well as open-world knowledge, may be necessary to improve results in the task. Furthermore, fact checking approaches should benefit from incorporating argumentation principles, defined as a "kind of discourse through which knowledge claims are individually and collaboratively constructed and evaluated" based on evidence (Sethi, 2017).This is a very interesting open research field.

As a final result, this task should be able to classify a claim based on its veracity as well as generate a report with the reasoning for the decision of the veracity value. Generating this report is complex when using machine learning approaches, thereby representing a future challenge for researchers in this area.

4.4. Clickbait detection

Clickbait refers to content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page. Typically, it is spread on social media in the form of short teaser messages that may read like the following cited examples:

- Man tries to hug a wild lion, You won't believe what happens next!
- A school girl gave her lunch to a homeless. What he did next will leave you in tears!

- Supermodels apply these three simple tricks to look young. Click to know what they are.

Clickbaits work by exploiting the insatiable appetite of humans to indulge their curiosity. Curiosity tends to proceed in two basic steps. First, a situation reveals a painful gap in our knowledge (that's the headline), and then we feel an urge to fill this gap and ease that pain (that's the click). In automatic processing, the clickbait detection task is a classification task where the headline is assessed as to whether it is a clickbait.

Section 4.4.1 reports the datasets and competitions found in the literature regarding clickbait detection. Section 4.4.2 presents systems structured according to the approach used to resolve the task (machine learning, deep learning and hybrid approaches). Finally, section 4.4.3 provides a discussion of the status quo of the subtask and areas for improvement.

4.4.1. Resources and competitions

Due to the lack of resources, the main available datasets at this moment are provided by researchers in the field of clickbait detection.

One of these datasets is presented in the work of Chakraborty et al. (2016), which has an even distribution of 7,500 clickbait headlines and 7,500 non-clickbait headlines. The non-clickbait headlines in the dataset were sourced from Wikinews, and the clickbait headlines were sourced from BuzzFeed, Upworthy, ViralNova, Scoopwhoop and ViralStories.

Biyani et al. (2016) dataset comes from different news sites whose pages surfaced on the Yahoo homepage, and included the Huffington Post, New York Times, CBS, Associated Press, Forbes, etc. They collected 1349 clickbait and 2724 non-clickbait web-pages. Rony et al. (2017) presented a dataset of 32,000 headlines.

Finally, Potthast et al. (2018b) created The Webis Clickbait Corpus 2017 (Webis-Clickbait-17,[22] for short) that was preceded by the Webis Clickbait Corpus 2016 (Potthast et al., 2016) to study clickbait detection for the first time. The main difference of this last dataset and the previous presented ones is that since all the other datasets are binary annotated, the latter is graded in a 4-point Likert scale, with values: Not clickbaiting (0.0), Slightly clickbaiting (0.33), Considerably clickbaiting (0.66), Heavily clickbaiting (1.0).

[22]https://webis.de/data/webis-clickbait-17.html (accessed online 28 February, 2019)

Regarding competitions, Potthast et al. (2018a) also organized the Clickbait challenge 2017. This competition used the Webis-Clickbait-17 dataset and it was focused on the detection of clickbait posts in social media.[23] The best system obtained an F-measure of 68.3% and 86% in accuracy.

Table 12 presents a summary of the datasets addressed in this section for clickbait detection purposes. In this table, apart from the reference of the work, the type of text annotated and the size of the dataset are presented. It is important to mention that all datasets presented in this table were annotated on a binary annotation scale (clickbait or not clickbait), except the latter which uses the four-point scale mentioned before.

Work	Source	Size
Biyani et al. (2016)	News articles	4,073 headlines
Chakraborty et al. (2016)	News articles	15,000 headlines
Potthast et al. (2016)	Twitter	2,992 tweets
Rony et al. (2017)	News articles	32,000 headlines
Webis Clickbait Corpus(Potthast et al., 2018b)	Twitter	38,517 tweets

Table 12: Summary of available resources for clickbait detection, according to the source, indicating the size reported

4.4.2. Systems

Research dealing with Clickbait detection are mainly applying machine learning approaches, but other approaches will also be presented.

Machine Learning approaches. Chakraborty et al. (2016) presents a system to automatically detect clickbaits, alerting the user of the possibility of being baited by a headline. After a detailed linguistic analysis of their own corpus of 15,000 headlines, they decided on a set of semantic and syntactic nuances to detect the clickbaits. These included the following: sentence structure; stop words; hyperbolic and common phrases; sentence subjects; and, determiners and possessives. Using these features, they perform a classification with three different ML methods: Support Vector Machines (SVM), Decision Trees (DT) and Random Forest (RF). The best classification method for their experiments was SVM, obtaining an F-measure of 93% and ROC-AUC of 97%.

[23] https://www.clickbait-challenge.org/ (accessed online 28 February, 2019)

The work presented at Biyani et al. (2016) applies machine learning to resolve the detection problem, more specifically Gradient Boosted Decision Trees (GBDT) that uses a set of content and similarity features, but with the novelty of taking into account features related to the informality of the text and the forward-reference to a concept/discourse/entity mentioned in the article. Informality is quite common in these types of articles and forward-reference features refers to the use of demonstratives or pronouns in the headline text to create a tease or information gap between the headline and the body text spurring curiosity among readers.

The work presented by Potthast et al. (2016) applied three different classifying machine learning methods: Logistic Regression (LR), Naive Bayes (NB) and Random Forest (RF). Their model is based on 215 features, divided into three main categories: (1) the teaser message, (2) the linked web page, and (3) meta information. Same authors (Potthast et al., 2018a) organized The Clickbait Challenge 2017 aforementioned and participated in the challenge using their system with the same feature set but replacing the random forest classifier with a ridge regression algorithm.

Finally, Bourgonje et al. (2017) system is based on simple, lemmatisation-based n-gram matching for the binary classification of "related" vs. "unrelated" headline/article pairs, obtaining the best results using a setup where the more fine-grained classification of the "related" pairs (into "agree", "disagree", "discuss") is carried out using a Logistic Regression classifier at first, and then three binary classifiers with slightly different training procedures for the cases where the first classifier lacked confidence. They experimented with FNC1 dataset and obtained and accuracy score of 89.59%.

Deep Learning approaches. Chakraborty et al. (2016) work is directly comparable to Anand et al. (2017), since they are using the same corpus created by (Chakraborty et al., 2016) for their evaluation. However, Anand is using deep learning to deal with the problem. More specifically their architecture is based on Recurrent Neural Networks, relying on distributed word representations learned from large corpora, and character embeddings learned via Convolutional Neural Networks. This approach surpasses the Chakraborty work, obtaining an F-score of 98% and ROC-AUC of 99%.

Rony et al. (2017) presents an approach using their own developed detection model that uses distributed word-embeddings learned from a large corpus. The system was built on their own corpus, created from news headlines and their contents, obtaining 98% of F-measure and the same value for

ROC-AUC.

Zingel system by Zhou (2017) was the best-performing approach of the Clickbait Challenge 2017 (Acc of 85.6%). It employs a neural network architecture with bidirectional gated recurrent units (biGRU) and a self-attention mechanism to assess clickbait strength (namely, the mean of the clickbait annotations for a tweet). As input to the network, only the teaser text is considered, which is represented as a sequence of word embeddings that have been pre-trained on Wikipedia using Glove (but then updated during training).

Hybrid approaches. The work presented by Chen et al. (2015) is a hybrid combination of methods to identify different clickbait cue types. To identify lexical and semantic patterns, they applied Support Vector Machines, Naive Bayes and Frequency analysis. Regarding syntax and pragmatics, they used Probability Context Free Grammar (PCFG) and Neural Network Analysis. They also performed image detection and analysis of the caption, as well as web traffic and metadata analysis for capturing user behaviour. In this case, no evaluation data was provided.

. In Table 13, a summary of the studies regarding clickbait detection are presented.

Work	Approach	Resource	Score
Anand et al. (2017)	DL(RNN+CNN)	Chakraborty et al.	0.980 (F1)
Biyani et al. (2016)	ML(GBDT)	Biyani	0.732 (F1)
Bourgonje et al. (2017)	ML(LR)	Fake News Challenge	0.8959 (FNC1)
Chakraborty et al. (2016)	ML(SVM)	Chakraborty	0.930 (F1)
Chen et al. (2015)	Hybrid	-	-
Potthast et al. (2016)	ML(LR/NB/RF)	Twitter Tweets	0.760 (F1)
Potthast et al. (2018a)	ML(Ridge Regr.)	Webis-Clickbait-17	0.552 (F1)
Rony et al. (2017)	DL(Word Embeddings)	Rony	0.983 (F1)
Zhou (2017)	DL(biGRU)	Webis-Clickbait-17	0.683 (F1)

Table 13: Summary of the analyzed studies on clickbait detection according to the computational approach, the resource used to train and the best score reported, including the metric inside brackets: F measure (F1) and Fake News Challenge-1 score (FNC1). An hyphen (-) is used where information is not provided

4.4.3. Discussion

Since the datasets used for evaluation are rather different, they cannot be directly compared from a performance point of view. The lack of gold stan-

dard datasets was a strong incentive for Clickbait Challenge organization[24] and thanks to this, more exhaustive and accurate comparisons can be made in future by continuing to hold such events and making available the datasets created from them. The conclusion that can be reached at this point from the review of studies presented is that applying deep learning techniques, such as Recurrent Neural Networks and Convolutional Neural Networks, improves the results compared to those obtained from classical machine learning approaches. This conclusion can be reached by comparing the systems that have been evaluated with the same dataset. For example, Chakraborty et al. (2016) and Anand et al. (2017) have used the same dataset to evaluate their systems, and as can be seen, the results obtained by Anand, who uses deep learning, are better than those obtained by Chakraborty, who applies machine learning (Improvement of 5% in F-measure). The two systems presented in this review, that are evaluated with the corpus Webis Clickbait Corpus 2017 (dataset provided by Clickbait Challenge 2017), follow the same line. Zhou's system (Anand et al., 2017), based on deep learning improves the results presented by Potthast et al. (2018a), who applies ML by 13.1% for F-Measure. Very recent research, such as TI-CNN (Kim, 2014) also use deep learning, and more specifically Convolutional Neural Networks in order to extract a set of latent features for capturing explicit and hidden text and image patterns appearing in fake news. However, at this moment, the differences in performance are not so great between ML and DL approaches. Furthermore, it is noteworthy that the results are much better for systems that use their own dataset, but less so when using a larger corpus such as the Clickbait Challenge, where none of the participating systems was above 70% in F-measure. This facts shows that the datasets used in some studies show certain deficiencies for the task and the analysis of their results must take this into consideration. In addition, the results obtained by the system using the dataset of the Clickbait Challenge, which is a external dataset, indicates there is still much room for improvement in this research area.

4.5. Credibility

This section focuses on the credibility of online information, which may present features that raise concerns about trustworthiness in terms of: i) the media; ii) the information source; and, iii) the message. A notable charac-

[24]https://www.clickbait-challenge.org/ (accessed online 28 February, 2019)

teristic for much online information is the relative absence of professional editors. While traditional media (newspapers, television, magazines, books, etc.) are subject to expert supervision of content and editorial review, web-based information does not always go through this process of quality control. Because of the presence of inaccurate, biased and false information online, assessing credibility is a major concern for today's society.

The following paragraphs present the core elements identified in the literature for: credibility assessment (Section 4.5.1); the process of data gathering and corpus development (Section 4.5.2); and, system development to automatically identify credibility (Section 4.5.3). Finally, a discussion section that summarizes and contrasts systems is presented.

4.5.1. Features

The study of media credibility in social sciences, psychology and marketing disciplines has a long history, and only recently has the identification of credibility been approached from a computational perspective. Different studies in credibility assessment have been focused on identifying the core features that humans judge to be key to determining the trustworthiness of the media, information source, and the message (Metzger et al., 2003).

Researchers have addressed the credibility problem in three different types of media: i) *online news* (Borah, 2014) (Howe & Teufel, 2014), ii) *blogs* (Gunter et al., 2009) (Matheson, 2004) (Rubin & Liddy, 2006) (Finn & de Ziga, 2011)(Johnson et al., 2007)(Johnson & Kaye, 2009) and, iii) *social networks* (Johnson & Kaye, 2014, 2015). The results showed that all traditional media were rated more credible than social media sites.

Most of the studies dealing with credibility in social networks chose Twitter as their focus of study. In this network, content and non-content features have been analyzed for credibility assessment. Content features were the focus of Alrubaian et al. (2017a), Alrubaian et al. (2017a) and Shao et al. (2018). The first two studies proposed a new reputation metric in Twitter, combining analysis of the user's reputation on a given topic within the social network, as well as a measure of the user's sentiment to identify topically relevant and credible sources of information. Shao et al. (2018) found evidence that social bots play a disproportionate role in spreading articles from low-credibility sources: bots amplify such content in the early spreading moments, before an article goes viral, targeting users with many followers through replies and mentions. Also non-content features have been the focus of different studies: Kang et al. (2015) identified that metadata and image

type elements are, in general, the strongest influencing factors in credibility assessments; Lin et al. (2016) identified that authority cues were the most credible, and that the presence of retweets reduced perceptions of source credibility compared to conditions with no retweets; Westerman et al. (2014) found that recency of tweets impacts source credibility, although this relationship was mediated by cognitive elaboration; Sandy et al. (2017) showed that verbal rather than non-verbal cues had more influence on participant judgments; and finally, in Shariff et al. (2017) the authors found that a reader's educational background, geo-location and news attributes (e.g. writing style) in tweets have significant correlation with credibility perception. Also non-content features have been the focus of different studies: Kang et al. (2015) identified that metadata and image type elements are, in general, the strongest influencing factors in credibility assessments; Lin et al. (2016) identified that authority cues were the most credible, and that the presence of retweets reduced perceptions of source credibility compared to conditions with no retweets; Westerman et al. (2014) found that the recency of tweets impacts source credibility, although this relationship was mediated by cognitive elaboration; Sandy et al. (2017) showed that verbal rather than non-verbal cues had more influence on participant judgments; and finally, in Shariff et al. (2017), the authors found that a reader's educational background, geo-location and news attributes (e.g. writing style) in tweets have significant correlation with credibility perception.

In addition to Twitter, other social media platforms have caught the attention of researchers. The study in Li & Suh (2015), centered on Facebook, identified five factors as the core elements of credibility assessment: medium dependency, interactivity, transparency, argument strength, and information quality.

Some of the features presented in this section to approach the credibility problem are difficult to obtain in an automatic way such as identifying the attractiveness, intelligence, and transparency of a source of information. Other credibility features can be more easily identified and included in an automatic system, such as number of mentions and retweets in Twitter, existence of hypertext references, use of multimedia content, and identification of message length. Finally, other features could be obtained using NLP tools, such as textual coherence (Abdolahi & Zahedi, 2016) and objectivity of the information (Wiebe & Riloff, 2005).

4.5.2. Resources

One of the main problems of evaluating credibility is the lack of publicly available standard datasets. Different approaches have been identified in the studies analyzed for gathering the datasets required in their experiments: crowdsourcing platforms, university students, web-based surveys, and web crawlers. Table 14 summarises the studies analyzed, identifying the sources, methodology and size of the datasets gathered.

Amazon Mechanical Turk[25] is an example of a *crowdsourcing platform* used to build credibility corpus. This platform has been used in different studies to assess the credibility of social network users. Such is the case of Kang et al. (2015) (193 Twitter and Reddit users) and Johnson & Kaye (2015) (1,267 Twitter and Facebook users). In Gupta et al. (2014) the authors used a different crowdsourcing platform, CrowdFlower,[26] to obtain 1,500 Twitter users for their study. This platform was also used by Shariff et al. (2017) (754 Twitter users).

University students are also a common source of information for research studies in user credibility. Such is the case of Armstrong & McAdams (2009) (1372 blog users), Borah (2014) (550 online newspapers users), Sandy et al. (2017) (24 Twitter users) and Westerman et al. (2014) (181 Twitter users).

Web-based surveys are another tool used for data gathering, used in several user credibility studies: Rickman et al. (2014) (285 blogs users), Kang (2010) (41 blog users), Li & Suh (2015) (135 Facebook users), Howe & Teufel (2014) (257 online news users), Finn & de Ziga (2011) (1,159 blog users), Johnson et al. (2007) and Johnson & Kaye (2009) (1,399 blog users), Johnson & Kaye (2014) (4,241 social networks users) and Lin et al. (2016) (696 Twitter users).

The last approach to data gathering consists of using *web crawlers* to leverage existing labelled datasets in the Web. This approach is usually used in studies developing machine learning based systems for credibility assessment, which require large datasets to train the models. Different websites and social networks have been exploited for this purpose. The most salient ones are mentioned in the remainder of this section.

Twitter was the focus of the work presented by Alrubaian et al. (2017b,a) (2,977,682 tweets), Abu-Salih et al. (2018) (2,810,362 tweets and 7,401 users),

[25]https://www.mturk.com/ (accesed online 28 February, 2019).
[26]http://www.crowdflower.com/ (accesed online 28 February, 2019).

Shao et al. (2018) (13,617,425 tweets linked to known low-credibility sources and 1,133,674 tweets linked to fact-checking sources) and Middleton (2015) (5,008 real and 7,032 fake tweets posted by 4,756 and 6,769 unique users respectively). Mitra & Gilbert (2015) presented CREDBANK, the first known attempt to provide a standard corpus with credibility judgments annotation. The corpus comprised more than 60 million tweets grouped into 1,049 real-world events. Another platform exploited by this approach is Sina Weibo,[27] the leading microblog service in China, used by Jin et al. (2014) (32 fake news items from fake news rank lists and 135 true news items from bot news), Jin et al. (2016) (73 fake news and 73 real news items) and Liu et al. (2016) (630,363 posts containing rumors and non rumors from 321,246 users).

Review sites have also been explored in several studies. Yelp,[28] which offers crowdsourced reviews about local businesses (e.g. restaurants and hairdressers), was used by Fontanarava et al. (2017) (134,724 recommended and 21,988 not recommended reviews) and Viviani & Pasi (2017) (140,000 recommended and 20,000 non recommended reviews). Seth et al. (2015) tested their model on a dataset of ratings obtained from digg.com, a website that allows users to submit links to news articles or blogs, which are called stories in the terminology used by the website. The dataset consisted on 85 stories with ratings by 27 users.

Fact-checking sites have also been used to obtain large datasets. Mukherjee & Weikum (2015) collected 18,500 news articles from NewsTrust, a news community (now offline) with available ground-truth ratings for credibility analysis of news articles. Popat et al. (2017) performed case studies of two real datasets: 133,272 web articles from Snopes.com and from Wikipedia pages.

Finally, online news sites have been the chosen source of information in several studies: Bountouridis et al. (2018) (85,405 news articles from major U.S. outlets) and Horne et al. (2018) (136,000 politics news articles from different websites and RSS feeds).

4.5.3. Systems

In the literature, the task of automatically assessing the credibility of online information has been mainly tackled as a classification problem. In

[27] https://www.weibo.com/ (accesed online 28 February, 2019).
[28] https://www.yelp.com/ (accesed online 28 February, 2019).

Work	Source	Size
Crowdsourcing platforms		
Gupta et al. (2014)	Twitter	1,500 users
Johnson & Kaye (2015)	Twitter and Facebook	1,267 users
Kang et al. (2015)	Twitter and Reddit	193 users
Shariff et al. (2017)	Twitter	754 users
University students		
Armstrong & McAdams (2009)	Blogs	1372 users
Borah (2014)	Online newspapers	550 users
Sandy et al. (2017)	Twitter	24 users
Westerman et al. (2014)	Twitter	181 users
Web-based surveys		
Rickman et al. (2014)	Blogs	285 users
Finn & de Ziga (2011)	Blogs	1,159 users
Howe & Teufel (2014)	Online news	257 users
Johnson & Kaye (2009)	Blogs	1,399 users
Johnson & Kaye (2014)	Social networks	4,241 users
Kang (2010)	Blogs	41 users
Li & Suh (2015)	Facebook	135 users
Lin et al. (2016)	Twitter	696 users
Web crawlers		
Abu-Salih et al. (2018)	Twitter	2,810,362 tweets and 7,401 users
Alrubaian et al. (2017b)	Twitter	2,977,682 tweets
Bountouridis et al. (2018)	Online news	85,405 articles
Fontanarava et al. (2017)	Yelp	156,712 reviews
Horne et al. (2018)	Websites	136,000 articles
Jin et al. (2014)	Sina Weibo	167 articles
Jin et al. (2016)	Sina Weibo	146 articles
Liu et al. (2016)	Sina Weibo	630,363 tweets and 321,246 users
Middleton (2015)	Twitter	12,040 tweets and 11,525 users
Mitra & Gilbert (2015)	Twitter	60,000,000 tweets
Mukherjee & Weikum (2015)	NewsTrust	18,500 articles
Popat et al. (2017)	Snopes	133,272 articles
	Wikipedia	100 hoaxes and 57 people pages
Seth et al. (2015)	digg.com	85 stories and 27 users
Shao et al. (2018)	Online news	389,569 articles
	Fact-checking sources	15,053 articles
	Twitter	14,751,099 tweets
Viviani & Pasi (2017)	Yelp	160,000 reviews

Table 14: Summary of corpus grouped by the approach adopted: crowdsourcing platforms, university students, web-based surveys and web crawlers.

this approach, machine learning techniques are used to categorize information (e.g. a blog or a post in Twitter) into truthful or fake depending on multiple kinds of textual and non-textual features. Table 15 summarizes the systems identified in this review, grouped by the type of media analyzed: online news, blogs and social networks.

Regarding *online news*, Bountouridis et al. (2018) created an interactive interface providing an additional information layer that is applied to an article's original textual content. This interface shows those pieces of information that are cross-referenced and thus, in the author's opinion, more likely to be credible. Horne et al. (2018) introduced an open source toolkit to explore the credibility of news articles using content-based markers of reliability and bias, training a Random Forest classifier on news sources. The work in Mukherjee & Weikum (2015) focused on identifying credibility for news communities. Their method extended Computational Random Fields (CRF) to identify highly credible news articles, trustworthy news sources and expert users. The research presented in Popat et al. (2017) focused on assessing the credibility of emerging claims with sparse presence in web-sources. The approach was based on distant supervision and CRF, exploiting the interaction taking place between various factors such us source reliability, stance over time and article objectivity.

Other systems have been developed to identify *blog* credibility. Juffinger et al. (2009) derived a credibility ranking function, based on the cosine similarity of vectors representing verified news articles and blog entries. Weerkamp & de Rijke (2008) proposed a blog post retrieval system based on language modeling, which incorporated textual credibility indicators at post level and blog level. The indicators used were based on the work by Rubin & Liddy (2006), keeping those that could be reliably estimated with state-of-the-art language technologies. The authors extended this work in Weerkamp & de Maarten Rijke (2012) for further exploring the impact of credibility-inspired indicators on the task of blog post retrieval.

Among the studied media, *social networks* have attracted the most attention from scholars to date. These studies have tried to address both content and user credibility. The credibility of users was the focus in Abu-Salih et al. (2018), which proposed a fine-grained credibility analysis framework to determine highly trustworthy users in specific domains. The proposal incorporates semantic analysis (e.g. sentiment identification) and temporal factors, obtaining as an output a ranked list of users with a corresponding credibility value for each specific domain.

Content credibility in Twitter has been the focus of Gupta et al. (2014). The authors faced the challenge of real-time credibility assessment on Twitter following a semi-supervised learning to rank approach, using SVM-rank. They used an extensive set of 45 features (including tweet meta-data, content, author, network, and links) to determine a credibility score. Middleton (2015) presented a semi-automated approach to trust and credibility analysis of tweets referencing suspicious images and videos. The authors used NLP to extract evidence from tweets in the form of fake and genuine claims attributed to trusted and untrusted sources. Sina Weibo has also been the focus of different studies. Wu et al. (2016) established a credibility evaluation platform that collects a database of rumors that have been verified on Sina Weibo and automatically evaluated the information which is generated by users on social media but has not been verified. The approach uses logistic regression to classify information into rumors and non-rumors. Liu et al. (2016) developed a method to learn representations of information credibility on Sina Weibo. Latent representations were learned from user credibility, behavior types, temporal properties and comment attitudes. Jin et al. (2014) presented a system to evaluate news credibility, based on a hierarchical propagation model. They built a three-layer credibility network consisting of event, sub-events and messages representing news from different scales. After linking these elements with their semantic and social associations, the credibility value of each element was propagated on this network to achieve the final evaluation result. The same authors applied this approach in Jin et al. (2016), although in this case they took advantage of "wisdom of crowds" to improve news verification by mining conflicting viewpoints. In a different evaluation domain, the authors of Fontanarava et al. (2017) analyzed the credibility of review comments in Yelp.com, considering separately textual (bag of words) and non-textual (e.g. number of friends of the user) features. They used an ensemble method to combine the results produced by two text classifiers (Logistic Regression and Recurrent Neural Networks) and another one using non-textual features (Random Forest).

Finally, there are systems that have used credibility as a supporting tool for a different task. In Sarna & Bhatia (2017), the authors proposed a system to identify cyber-bulling. They used a machine learning classifier to identify bullying and non-bullying messages, and then applied a rule-based system to decide the credibility of the user considering the type of messages sent. Seth et al. (2015) developed a recommender system that incorporated the credibility of messages to enhance the performance of collaborative filtering

recommendations. Their methodology rested on a Bayesian network based credibility model. Also related to recommender systems, the work in Zhou et al. (2017) proposed a social network user credibility method for generating reliable recommended items lists. The reviewer credibility was calculated by exploiting the correlation between reliable reviews and their maker. In the context of opinion spam detection in review sites, Viviani & Pasi (2017) proposed a system based on Multi-Criteria Decision Making, which included two aggregation schemes that considered in the credibility assessment process, the unequal importance of features.

Work	Approach	Resource	Score
Online news			
Bountouridis et al. (2018)	Genetic Algorithm	Online news	-
Horne et al. (2018)	Random Forest	Websites	0.89 (AUC)
Mukherjee & Weikum (2015)	CRF	NewsTrust	0.33 (MSE)
Popat et al. (2017)	CRF	Twitter	0.8139 (Acc)
Blogs			
Juffinger et al. (2009)	Cosine similarity	APA news corpus	0.83 (Prec)
Weerkamp & de Rijke (2008)	Language modeling	TREC Blog06	0.3550 (MAP)
Weerkamp & de Maarten Rijke (2012)	Language modeling	TREC Blog06	0.3893 (MAP)
Social networks			
Abu-Salih et al. (2018)	Matrix similarity	Twitter	0.85 (Prec)
Fontanarava et al. (2017)	RNN + Log. Regr.	Yelp.com	0.911 (Prec)
Gupta et al. (2014)	SVM-rank	Twitter	0.4014 (Prec)
Jin et al. (2016)	Hierarchical prop.	Sina Weibo	0.853 (F1)
Liu et al. (2016)	Latent representation	Sina Weibo	0.831 (Prec)
Middleton (2015)	Rule based	Twitter	0.83 (F1)
Sarna & Bhatia (2017)	SVM	Twitter	0.2238 (Error)
Seth et al. (2015)	Bayesian Network	Twitter	0.156 (MCC)
Viviani & Pasi (2017)	Multi-Criteria Decision	Twitter	0.82 (F1)
Wu et al. (2016)	Log. Regression	Sina Weibo	-
Zhou et al. (2017)	Matrix similarity	Amazon.com	0.0836 (MAP)

Table 15: Summary of systems organised by media analyzed: online news, blogs and social networks. *Score* represents the best score obtained by the system, including the metric inside brackets: precision (Prec), area under the curve (AUC), F measure (F1), mean squared error (MSE), accuracy (Acc), classification error (Error), Matthews correlation coefficient (MCC) and mean average precision (MAP). An hyphen (-) is used where information is not provided or no single best score can be reported.

4.5.4. Discussion

One of the problems presented by the existing research on credibility in online news, blogs, and social networks is the limited reliability of metrics partly due to the use of general credibility models that have been developed to assess the trustworthiness of traditional news media Kang (2010). Another problem is that these studies sometimes rely on surveys and focus groups in order to validate their proposals, and the profile of the surveyed users may introduce a bias in the results. Such is the case of user studies supported by university students, as described in section 4.5.2, a young population whose perception of credibility of online media largely differs from that of an older population, especially when it comes to judging social networks.

It is noteworthy that the existing datasets lack gold standards that serve to train and test systems for automatic credibility assessment. All the studies presented in this section rely on their own datasets to evaluate the systems, independently of the type of media studied. The only known attempt to gather and share a corpus to help in credibility assessment was developed by Mitra & Gilbert (2015), although none of the analyzed systems in section 4.5.3 used it in their experiments.

This absence of common datasets makes it difficult to compare the performance of different systems given the disparity in the size of the corpus employed. For instance, in the case of credibility systems in Twitter, the figures vary from a few thousand tweets Middleton (2015) to several million tweets Alrubaian et al. (2017b). Another problem is the use of different metrics in these studies, such as precision, MAP or F-measure for systems that perform binary classification (credible and non-credible) to AUC and MSE for systems that provide a continuous value of trustworthiness. The best performing system to date was developed by Fontanarava et al. (2017), which reported 91.1% precision in classifying reviews from Yelp.com. However, due to the problems mentioned before, it is not possible to extrapolate these results to other corpora or types of media.

5. Open issues on Fake News Detection

Following a thorough analysis of the previous research work, a review of their main conclusions is presented, which highlights the gaps in the state of the art for tackling the challenges of fake news detection, that need to be accomplished in the near future.

Gold standard resources. The lack of gold standard resources and datasets, as well as annotation guidelines is one of the most addressed problems in fake news detection, and this contributes to making the evaluation and training tasks more difficult to accomplish. This is the case when machine learning techniques, and in general, artificial intelligence procedures are applied. Given the novelty of this research field, the number of related competitions in which corpora is usually released is still limited. The lack of a standard annotation is also a problem, which results in most researchers creating their own dataset. Thus, two main directions in this issue are required: i) Standard data formats, so that any new automated tool can work with any known source, and ii) Open and shared evaluation competitions to compare systems and know exactly how they work. Moreover, all the datasets provided at this moment are mainly developed in English. However, as fake news is a global problem, more resources in different languages are needed to cover the problem on world-wide scale.

Fake News Rich Models. The research work reviewed in this paper shows different Artificial Intelligence strategies, mainly machine learning, but there is a lack of in-depth modelling of the problem. Generally, the research reviewed is based on the most obvious characteristics of the text (linguistic-based cues) but they omit the intrinsically sociological and psychological characteristics of the process of deception. In that sense, modelling should go a step further, enabling the discovery of extra-linguistic relationships that lead to the interpretation of lies. In addition, so far, the work done is not able to justify the reason for the inferences obtained, mainly due to the hidden deductive process of machine learning techniques, especially as the complexity of the problem increases. Changing people's stance on the veracity of information involves not only a classification of true or false values, but also a justified reasoning for the choice. Therefore, completing the fake news detection challenge requires a justification of the classification result, and that has not been addressed so far in the research.

Further linguistic patterns and cues. Previous surveys have pointed out the importance of determining linguistic patterns and cues to detect deception in written languages. However, further examination of the linguistic patterns and cues must be performed. Three types of linguistic information is required: i) Additional syntactic features: It is necessary to evaluate whether additional language features, such as grammatical dependencies or named

entity labels, would improve the process, ii) Additional semantic features: There is a lot to be done in this area, since establishing semantic relations between entities (in this and other documents previously processed) are necessary to infer the veracity of news, and iii) Pragmatic information: determining spatio-temporal frames are crucial for fake news detection, as well as the context of news, including the speaker, the topic, the audience, and so on. However, due to the complexity of pragmatic information processing, its inclusion in the fake news detection process is still an open challenge. It is also necessary to consider that the production mechanism of fake news is totally dynamic, and conditioned by the evolution of the technological platforms as well as by the behaviour patterns of the deceivers.

Suitable Technologies. In general, the reviewed studies show a tendency to apply trending artificial intelligence techniques, ranging from classical machine learning to complex deep learning strategies, with the latter showing very promising results when applied to other AI problems. However, the classical techniques as they are being addressed so far fail to overcome a certain threshold of effectiveness and appear to have peaked. Meanwhile, according to the review, the promising techniques of deep learning do not demonstrate outstanding improvement despite the computational cost and resources that these techniques require. This makes it necessary to address specific techniques that are adapted to the problem of written language, such as hybrid combinations that allow a fine tuning of the problem. Some previously mentioned authors are already considering the importance of exploiting information extracted from social networks. All the information that can be sourced from social networks regarding specific news or topics, including social network structures, user behaviour, impact in social networks, and viralization is adding additional and crucial information to determine whether news is fake. Very recently, there are interesting proposals to be assessed that involve collecting facts about news events using a crowdsourced knowledge graph, which is dynamically updated by local and well-informed people. This timely information could be used to compare against that extracted from news articles, thereby helping to detect fake news (Zhou et al., 2019).

Emotion-driven news. Automatically determining the emotional burden of news and how this affects a specific profile of users is a challenging task, considering that one of the main features of fake news is the highly emotional tone deployed. There are social science researchers, like Vosoughi

et al. (2018), that establish a relationship between the different emotions that fake and true news generate among users. For example, they explain that false stories inspired fear, disgust and surprise in replies, whereas true stories inspired anticipation, sadness, joy and trust. Automatically detecting these emotions in written texts may help fake news detection. The relationship between virality and emotions is also being studied by psychologists (Guerini & Staiano, 2015) and it could be a promising future work for NLP. Furthermore, using emotions detection in online social collaborative argumentation applied to verify the statements contained in news and opinion articles is a very interesting research line (Sethi & Rangaraju, 2018).

Classical unresolved NLP problems. Dealing with classical problems like paraphrasing or irony and satire is still necessary since these elements are part of fake news.

Combining subtasks for a complete Fake News Detection architecture. One of the most challenging open issues would be the development of a methodology capable of combining not only the tasks presented in this review and which are currently being dealt with, but also the inclusion of the open problems previously presented here. Finding the most appropriate interrelation between subtasks, as well as the best approaches to deal with each one of them is a great challenge due to the complexity involved in achieving automatic fake news detection.

6. Conclusions

This paper presents an extensive and systematic review of the different areas that are involved in the automatic detection of fake news from the perspective of Natural Language Processing (NLP). For this purpose, the problem has been defined as well as a classification of the different scenarios where it is likely to occur. After applying a divide-and-conquer methodology, different research lines and tasks have emerged from it, resulting in a thorough review of the state of the art. These lines include: Deception detection, stance detection, controversy, polarization, automated fact checking, clickbait detection, and credibility. For each of these lines, the techniques applied are presented as well as the most recent resources and competitions available to the research community. In this sense, it is important to point out that these technological research lines are complementary, since the tasks themselves often share common objectives and problems. Among these common

problems we highlight that all the subtasks presented are not fully addressed (simplifications of the overall issue are being addressed). Furthermore, the results reported in the studies are rather limited. There is still much room for improvement in all tasks. There is a significant lack of clear and reliable metrics to standardize tasks, as well as a lack of well-built (balanced and reliable) resources. In those subtasks that apply deep learning approaches (stance detection and clickbait), these approaches improve the results compared to those systems applying machine learning, but the differences are not significant at this moment, so it is essential to continue working on the definition of the most relevant characteristics (the construction of the language model for each task). Finally, a set of open challenges have been extracted from the different studies that should be addressed in the near future. These include the following: the creation of gold standard resources; the development of rich models on fake news; the incorporation of further linguistic patterns and cues; the selection of suitable technologies adapted to specific scenarios; the incorporation of emotion evaluation techniques associated with the production of fake news; and, the resolution of classic NLP problems, such as paraphrasing and irony. This would provide guidelines for a roadmap that includes the future challenges that could be addressed by the research community.

Acknowledgements

This research work has been partially funded by Generalitat Valenciana through project "SIIA: Tecnologias del lenguaje humano para una sociedad inclusiva, igualitaria, y accesible" with grant reference PROMETEU/2018/089, by the Spanish Government through project RTI2018-094653-B-C22: "MODELADO DEL COMPORTAMIENTO DE ENTIDADES DIGITALES MEDIANTE TECNOLOGIAS DEL LENGUAJE HUMANO", as well as by the project "Analisis de Sentimientos Aplicado a la Prevencion del Suicidio en las Redes Sociales (ASAP)" funded by Ayudas Fundacion BBVA a equipos de investigacion cientifica.

References

Abdolahi, M., & Zahedi, M. (2016). An overview on text coherence methods. In *2016 Eighth International Conference on Information and Knowledge Technology (IKT)* (pp. 1–5). IEEE.

1861 Abu-Salih, B., Wongthongtham, P., Chan, K. Y., & Zhu, D. (2018). Cred-
1862 sat: Credibility ranking of users in big social data incorporating semantic
1863 analysis and temporal factor. *Journal of Information Science*, *44*, 1–22.

1864 Almela, A., Valencia-García, R., & Cantos, P. (2012). Seeing through de-
1865 ception: A computational approach to deceit detection in written commu-
1866 nication. In *Proceedings of the Workshop on Computational Approaches*
1867 *to Deception Detection* EACL 2012 (pp. 15–22). Stroudsburg, PA, USA:
1868 Association for Computational Linguistics.

1869 Alrubaian, M., Al-Qurishi, M., Al-Rakhami, M., & Alamri, A. (2017a). A
1870 credibility assessment model for online social network content. In *From*
1871 *Social Data Mining and Analysis to Prediction and Community Detection*
1872 (pp. 61–77). Cham: Springer International Publishing.

1873 Alrubaian, M., Al-Qurishi, M., Al-Rakhami, M., Hassan, M. M., & Alamri,
1874 A. (2017b). Reputation-based credibility analysis of twitter social network
1875 users. *Concurrency and Computation: Practice and Experience*, *29*, 1–112.

1876 Amoros, M. (2018). *Fake News: La verdad de las noticias falsas*. Plataforma
1877 Actual.

1878 Anand, A., Chakraborty, T., & Park, N. (2017). We used neural networks to
1879 detect clickbaits: You won't believe what happened next! In J. M. Jose,
1880 C. Hauff, I. S. Altingövde, D. Song, D. Albakour, S. N. K. Watt, & J. Tait
1881 (Eds.), *Advances in Information Retrieval - 39th European Conference on*
1882 *IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*
1883 (pp. 541–547). volume 10193 of *Lecture Notes in Computer Science*.

1884 Armstrong, C. L., & McAdams, M. J. (2009). Blogs of information: How
1885 gender cues and individual motivations influence perceptions of credibility.
1886 *Journal of Computer-Mediated Communication*, *14*, 435–456.

1887 Ba, M. L., Berti-Equille, L., Shah, K., & Hammady, H. M. (2016). Vera:
1888 A platform for veracity estimation over web data. In *Proceedings of the*
1889 *25th International Conference Companion on World Wide Web* WWW '16
1890 Companion (pp. 159–162). Republic and Canton of Geneva, Switzerland:
1891 International World Wide Web Conferences Steering Committee.

1892 Babakar, M., Bakos, N., Daum, H., Mantzarlis, A., Seddah, D., Vlachos, A.,
1893 & Wardle, C. (2016). Fake news challenge - i.

Baird, S., Sibley, D., & Pan, Y. (2018). Talos targets disinformation with fake news challenge victory.

Banerjee, S., & Chua, Y. (2014). Applauses in hotel reviews: Genuine or deceptive? In *Proceedings of the 2014 Science and Information Conference* (pp. 938–94). IEEE.

Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Atanasova, P., Zaghouani, W., Kyuchukov, S., Da San Martino, G., & Nakov, P. (2018). Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims, task 2: Factuality. In L. Cappellato, N. Ferro, J.-Y. Nie, & L. Soulier (Eds.), *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum* CEUR Workshop Proceedings. Avignon, France: CEUR-WS.org.

Bastos, M. T., & Mercea, D. (2019). The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, *37*, 38–54.

Biyani, P., Tsioutsiouliklis, K., & Blackmer, J. (2016). "8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* AAAI'16 (pp. 94–100). AAAI Press.

Borah, P. (2014). The hyperlinked world: A look at how the interactions of news frames and hyperlinks influence news credibility and willingness to seek information. *Journal of Computer-Mediated Communication*, *19*, 576–590.

Bountouridis, D., Marrero, M., Tintarev, N., & Hauff, C. (2018). Explaining credibility in news articles using cross-referencing. In *Proceedings of the 1st International Workshop on ExplainAble Recommendation and Search (EARS 2018)*. MI, USA: Association for Computational Linguistics SIGIR.

Bourgonje, P., Moreno Schneider, J., & Rehm, G. (2017). From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism* (pp. 84–89). Association for Computational Linguistics.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in twitter during the 2016 US presidential election. *Nature Communications*, *10(1):7*.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 632–642). Association for Computational Linguistics.

Burgoon, J., Buller, D., Guerrero, L., Afifi, W., & Feldman, C. (1996). Interpersonal deception: Xii information management dimensions underlying deceptive and truthful messages. *Communication Monographs*, *63*, 52–69.

Burgoon, J. K., Blair, J. P., Qin, T., & Nunamaker, J. F., Jr. (2003). Detecting deception through linguistic analysis. In *Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics* ISI'03 (pp. 91–101). Berlin, Heidelberg: Springer-Verlag.

Cazalens, S., Lamarre, P., Leblay, J., Manolescu, I., & Tannier, X. (2018). A content management perspective on fact-checking. In *WWW (Companion Volume)* (pp. 565–574). Association for Computational Linguistics.

Chakraborty, A., Paranjape, B., Kakarla, S., & Ganguly, N. (2016). Stop clickbait: Detecting and preventing clickbaits in online news media. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* ASONAM '16 (pp. 9–16). Piscataway, NJ, USA: IEEE Press.

Chaudhry, A. K., Baker, D., & Thun-Hohenstein, P. (2017). Stance detection for the fake news challenge: identifying textual relationships with deep neural nets. *CS224n: Natural Language Processing with Deep Learning*, .

Chen, Y., & Chen, H. (2014). Opinion spam detection in web forum: A real case study. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 173–183).

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). Misleading online content: Recognizing clickbait as "false news". In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection* WMDD '15 (pp. 15–19). New York, NY, USA: Association for Computational Linguistics.

Ciampaglia, G. L., Mantzarlis, A., Maus, G., & Menczer, F. (2018). Research challenges of digital misinformation: Toward a trustworthy web. *AI Magazine*, *39*, 65–74.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. *PLoS One*, *10(6)*, 1–13.

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community* ASIST '15 (pp. 82:1–82:4). Silver Springs, MD, USA: American Society for Information Science.

Dale, R. (2017). Nlp in a post-truth world. *Natural Language Engineering*, *23*, 319–324.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., K., C., & H., C. (2003). Cues to deception. *Psychological Bulletin*, *129*, 74 – 118.

Derczynski, L., & Bontcheva, K. (2014). Pheme: Veracity in digital social networks. In I. Cantador, M. Chi, R. Farzan, & R. Jäschke (Eds.), *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization co-located with the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP2014), Aalborg, Denmark, July 7-11, 2014.*. CEUR-WS.org volume 1181 of *CEUR Workshop Proceedings*.

Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G., & Zubiaga, A. (2017). Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 69–76). Association for Computational Linguistics.

Dias, M., & Becker, K. (2016). Inf-ufrgs-opinion-mining at semeval-2016 task 6: Automatic generation of a training corpus for unsupervised identification of stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 378–383). Association for Computational Linguistics.

Dori-Hacohen, S., & Allan, J. (2015). Automated controversy detection on the web. In A. Hanbury, G. Kazai, A. Rauber, & N. Fuhr (Eds.), *Advances in Information Retrieval* (pp. 423–434). Cham: Springer International Publishing.

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., & Vrandečić, D. (2014). Introducing wikidata to the linked data web. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. A. Knoblock, D. Vrandečić, P. T. Groth, N. F. Noy, K. Janowicz, & C. A. Goble (Eds.), *Proceedings of the 13th International Semantic Web Conference (ISWC 2014)* (pp. 50–65). Springer volume 8796 of *LNCS*.

Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, *1*, 291–316.

Ferreira, W., & Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1163–1168). Association for Computational Linguistics.

Fillmore, C., Johnson, C., & Petruck, M. (2003). Background to framenet. *International Journal of Lexicography*, *16*, 235–250.

Finn, J., & de Ziga, H. G. (2011). Online credibility and community among blog users. *Proceedings of the American Society for Information Science and Technology*, *48*, 1–9.

Finn, S., Mustafaraj, E., & Metaxas, P. T. (2014). The co-retweeted network and its applications for measuring the perceived political polarization. In *Proceedings of the 10th International Conference on Web Information Systems and Technologies* (pp. 276–284). SciTePress.

Fitzpatrick, E., & Bachenko, J. (2012). Building a data collection for deception research. In *Proceedings of the Workshop on Computational Approaches to Deception Detection* (pp. 31–38). Avignon, France: Association for Computational Linguistics.

Fontanarava, J., Pasi, G., & Viviani, M. (2017). An ensemble method for the credibility assessment of user-generated content. In *Proceedings of the International Conference on Web Intelligence* WI '17 (pp. 863–868). New York, NY, USA: Association for Computational Linguistics.

Fornaciari, T., & Poesio, M. (2012). Decour: a corpus of deceptive statements in italian courts. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doan,

B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Francis, D., & Fact, F. (2017). Fast and furious fact check challenge 2016.

Fuller, C. M., Biros, D. P., & Wilson, R. L. (2009). Decision support for determining veracity via linguistic-based cues. *Decision Support Systems*, *46*, 695 – 703. Wireless in the Healthcare.

FullFact.org (2016). The state of automated factchecking (2016).

Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Quantifying controversy on social media. *Trans. Soc. Comput.*, *1*, 3:1–3:27.

George, J. F., Giordano, G., & Tilley, P. A. (2016). Website credibility and deceiver credibility: Expanding prominence-interpretation theory. *Computers in Human Behavior*, *54*, 83–93.

Gokhman, S., Hancock, J., Prabhu, P., Ott, M., & Cardie, C. (2012). In search of a gold standard in studies of deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection* EACL 2012 (pp. 23–30). Stroudsburg, PA, USA: Association for Computational Linguistics.

Guerini, M., & Staiano, J. (2015). Deep feelings: A massive cross-lingual study on the relation between emotions and virality. In *Proceedings of the 24th International Conference on World Wide Web* WWW '15 Companion (pp. 299–305). New York, NY, USA: Association for Computational Linguistics.

Gunter, B., Campbell, V., Touri, M., & Gibson, R. (2009). Blogs, news and credibility. *Aslib Proceedings*, *61*, 185–204.

Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014). Tweetcred: Real-time credibility assessment of content on twitter. In L. M. Aiello, & D. McFarland (Eds.), *Social Informatics* (pp. 228–243). Cham: Springer International Publishing.

Habernal, I., Wachsmuth, H., Gurevych, I., & Stein, B. (2018). The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT)* (pp. 1930–1940). New Orleans, LA, USA: Association for Computational Linguistics.

Hanselowski, A. (2018). The blog of team athene on the fake news challenge.

Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., & Gurevych, I. (2018). A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1859–1874). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Hassan, N., Arslan, F., Li, C., & Tremayne, M. (2017a). Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017* (pp. 1803–1812). Association for Computational Linguistics.

Hassan, N., Li, C., & Tremayne, M. (2015). Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* CIKM '15 (pp. 1835–1838). New York, NY, USA: Association for Computational Linguistics.

Hassan, N., Sultana, A., Wu, Y., Zhang, G., Li, C., Yang, J., & Yu, C. (2014). Data in, fact out: Automated monitoring of facts by factwatcher. *The Proceedings of the VLDB Endowment*, *7*, 1557–1560.

Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A. K., Sable, V., Li, C., & Tremayne, M. (2017b). Claimbuster: The first-ever end-to-end fact-checking system. *The Proceedings of the VLDB Endowment*, *10*, 1945–1948.

Hauch, V., Masip, J., Blandón-Gitlin, I., & Sporer, S. L. (2012). Linguistic cues to deception assessed by computer programs: A meta-analysis. In *Proceedings of the Workshop on Computational Approaches to Deception*

Detection EACL 2012 (pp. 1–4). Stroudsburg, PA, USA: Association for Computational Linguistics.

Hernández, D., Montes y Gómez, M., Rosso, P., & Guzman, R. (2015). Detecting positive and negative deceptive opinions using pu-learning. *Information Processing and Management*, *51*, 433–443.

Höfer, E., Akerhust, L., & Metzger, G. (1996). Reality monitoring: A chance for further development of cbca? In *Annual Meeting of the European Association on Psychology and Law*.

Hoffart, J., M. Suchanek, F., Berberich, K., & Weikum, G. (2013). Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, *194*, 28–61.

Hooper, V. (2018). Fake news and social media: The role of the receiver. In *ECSM 2018 5th European Conference on Social Media* (p. 62). Academic Conferences and publishing limited.

Horne, B. D., Dron, W., Khedr, S., & Adali, S. (2018). Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Companion Proceedings of the The Web Conference 2018* WWW '18 (pp. 235–238). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Howe, P., & Teufel, B. (2014). Native advertising and digital natives: The effects of age and advertisement format on news website credibility judgments. *ISOJ Journal*, *4*, 78–90.

Issenberg, S. (2013). *The Victory Lab: The Secret Science of Winning Campaigns*. Broadway Books.

Jang, M., Foley, J., Dori-Hacohen, S., & Allan, J. (2016). Probabilistic approaches to controversy detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* CIKM '16 (pp. 2069–2072). New York, NY, USA: Association for Computational Linguistics.

Jin, Z., Cao, J., Jiang, Y.-G., & Zhang, Y. (2014). News credibility evaluation on microblog with a hierarchical propagation model. In *Proceedings of the*

2014 IEEE International Conference on Data Mining ICDM '14 (pp. 230–239). Washington, DC, USA: IEEE Computer Society.

Jin, Z., Cao, J., Zhang, Y., & Luo, J. (2016). News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* AAAI'16 (pp. 2972–2978). AAAI Press.

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* WSDM '08 (pp. 219–230). New York, NY, USA: Association for Computational Linguistics.

Johnson, T. J., & Kaye, B. K. (2009). In blog we trust? deciphering credibility of components of the internet among politically interested internet users. *Computers in Human Behavior*, *25*, 175–182.

Johnson, T. J., & Kaye, B. K. (2014). Credibility of social network sites for political information among politically interested internet users. *Journal of Computer-Mediated Communication*, *19*, 957–974.

Johnson, T. J., & Kaye, B. K. (2015). Reasons to believe: Influence of credibility on motivations for using social networks. *Computers in Human Behavior*, *50*, 544–555.

Johnson, T. J., Kaye, B. K., Bichard., S. L., & Wong, W. J. (2007). Every blog has its day: Politically-interested internet users' perceptions of blog credibility. *Journal of Computer-Mediated Communication*, *13*, 100–122.

Juffinger, A., Granitzer, M., & Lex, E. (2009). Blog credibility ranking by exploiting verified content. In *Proceedings of the 3rd Workshop on Information Credibility on the Web* WICOW '09 (pp. 51–58). New York, NY, USA: Association for Computational Linguistics.

Kang, B., Hllerer, T., & O'Donovan, J. (2015). Believe it or not? analyzing information credibility in microblogs. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 611–616).

Kang, M. (2010). *Measuring Social Media Credibility: A Study on a Measure of Blog Credibility*. Technical Report S. I. Newhouse School of Public Communications, Syracuse University.

Karadzhov, G., Nakov, P., Màrquez, L., Barrón-Cedeño, A., & Koychev, I. (2017). Fully automated fact checking using external sources. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017* (pp. 344–353). INCOMA Ltd.

Kim, S., Chang, H., Lee, S., Yu, M., & Kang, J. (2015). Deep semantic frame-based deceptive opinion spam analysis. In *In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 1131–1140). Association for Computational Linguistics.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746–1751). Association for Computational Linguistics.

Krejzl, P., & Steinberger, J. (2016). Uwb at semeval-2016 task 6: stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 408–412).

Li, F., Huang, M., Yang, Y., & Zhu, X. (2011). Learning to identify review spam. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three* IJCAI'11 (pp. 2488–2493). AAAI Press.

Li, J., Ott, M., Cardie, C., & Hovy, E. (2014). Towards a general rule for identifying deceptive opinion spam. In *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 1566–1576). The Association for Computer Linguistics.

Li, R., & Suh, A. (2015). Factors influencing information credibility on social media platforms: Evidence from facebook pages. *Procedia Computer Science*, *72*, 314–328.

Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., & Lauw, H. W. (2010). Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM International Conference on Information and Knowledge*

Management CIKM '10 (pp. 939–948). New York, NY, USA: Association for Computational Linguistics.

Lin, X., Spence, P. R., & Lachlan, K. A. (2016). Social media and credibility indicators: The effect of influence cues. *Computers in Human Behavior*, *63*, 264–271.

Litvinova, O., Seredin, P., Sboev, A., & Romanchencko, O. (2016). Ruspersonality: A russian corpus for authorship profiling and deception detection. In *Proceedings of International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)* (pp. 1–7). IEEE.

Liu, Q., Wu, S., Yu, F., Wang, L., & Tan, T. (2016). Ice: Information credibility evaluation on social media via representation learning. *arXiv preprint arXiv:1609.09226*, .

Long, Y., Lu, Q., Xiang, R., Li, M., & Huang, C.-R. (2017). Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 252–256). Asian Federation of Natural Language Processing.

M. DePaulo, B., A. Kashy, D., E. Kirkendol, S., M. Wyer, M., & Epstein, J. (1996). Lying in everyday life. *Journal of personality and social psychology*, *70*, 979–95.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* IJCAI'16 (pp. 3818–3824). AAAI Press.

Manolescu, I. (2017). ContentCheck: Content Management Techniques and Tools for Fact-checking. *ERCIM News*, .

Matheson, D. (2004). Weblogs and the epistemology of the news: Some trends in online journalism. *New Media & Society*, *6*, 443–468.

McIntyre, L. (2018). *Post-Truth*. The MIT Press Essential Knowledge series. The MIT Press.

Metaxas, P. T., Finn, S., & Mustafaraj, E. (2015). Using twittertrails.com to investigate rumor propagation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work and Social Computing* CSCW'15 Companion (pp. 69–72). New York, NY, USA: Association for Computational Linguistics.

Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., & Mccann, R. M. (2003). Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Annals of the International Communication Association*, *27*, 293–335.

Middleton, S. (2015). Extracting attributed verification and debunking reports from social media: Mediaeval-2015 trust and credibility analysis of image and video. In *MediaEval 2015*. CEUR Workshop Proceedings.

Mihalcea, R., & Strapparava, C. (2009). The lie detector: explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference* (pp. 309–312). Association for Computational Linguistics.

Mitra, T., & Gilbert, E. (2015). Credbank: A large-scale social media corpus with associated credibility annotations. In *International AAAI Conference on Web and Social Media*.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 31–41).

Mukherjee, S., & Weikum, G. (2015). Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* CIKM '15 (pp. 353–362). New York, NY, USA: Association for Computational Linguistics.

Nakov, P., Màrquez, L., Barrón-Cedeño, A., Zaghouani, W., Elsayed, T., Suwaileh, R., & Gencheva, P. (2018). CLEF-2018 lab on automatic identification and verification of claims in political debates. In *Proceedings of the CLEF-2018*.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, *29*, 665 – 675.

Nie, Y., Chen, H., & Bansal, M. (2018). Combining fact extraction and verification with neural semantic matching networks. *CoRR*, *abs/1811.07039*, 1–10.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* HLT '11 (pp. 309–319). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pérez-Rosas, V., & Mihalcea, R. (2014). Cross-cultural deception detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)* (pp. 440–445).

Pérez-Rosas, V., & Mihalcea, R. (2015). Experiments in open domain deception detection. In *Proceedings of the 2015 Conference on Emprirical Methods in Natural Language Processing* (pp. 1120–1125). Association for Computational Linguistics.

Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2017). Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion* WWW '17 Companion (pp. 1003–1012). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Potthast, M., Gollub, T., Hagen, M., & Stein, B. (2018a). The clickbait challenge 2017: Towards a regression model for clickbait strength. *CoRR*, *abs/1812.10847*, 1–6.

Potthast, M., Gollub, T., Komlossy, K., Schuster, S., Wiegmann, M., Garces Fernandez, E. P., Hagen, M., & Stein, B. (2018b). Crowdsourcing a large corpus of clickbait on twitter. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1498–1507). Association for Computational Linguistics.

Potthast, M., Köpsel, S., Stein, B., & Hagen, M. (2016). Clickbait detection. In N. Ferro, F. Crestani, M. Moens, J. Mothe, F. Silvestri, G. M. D. Nunzio, C. Hauff, & G. Silvello (Eds.), *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings* (pp. 810–817). Springer volume 9626 of *Lecture Notes in Computer Science*.

Rakholia, N., & Bhargava, S. (2016). *Is it true?–Deep Learning for Stance Detection in News*. Technical Report Stanford University, California, USA.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2931–2937). Association for Computational Linguistics.

Ren, Y., & Ji, D. (2017). Neural networks for deceptive opinion spam detection: An empirical study. *Informtion Sciences*, *385-386*, 213–224.

Rickman, C. T., R., S. M., & Wisuk, K. (2014). Credibility in the blogosphere: A study of measurement and influence of wine blogs as an information source. *Journal of Consumer Behaviour*, *14*, 71–91.

Riedel, B., Augenstein, I., Spithourakis, G. P., & Riedel, S. (2017). A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *CoRR*, *abs/1707.03264*.

Rony, M. M. U., Hassan, N., & Yousuf, M. (2017). Diving deep into clickbaits: Who use them to what extents in which topics with what effects? In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* ASONAM '17 (pp. 232–239). New York, NY, USA: Association for Computational Linguistics.

Rubin, V., Conroy, N., & Chen, Y. (2015). Towards news verification: Deception detection methods for news discourse. In *Proceedings of the Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, Hawaii International Conference on System Sciences* HICSS48 (pp. 1–11).

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings*

of the Second Workshop on Computational Approaches to Deception Detection (pp. 7–17). Association for Computational Linguistics.

Rubin, V. L., & Liddy, E. D. (2006). Assessing credibility of weblogs. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium* (pp. 187–190).

Rubin, V. L., & Lukoianova, T. (2015). Truth and deception at the rhetorical structure level. *JASIST*, *66*, 905–917.

Ruby, C., & Brigham, J. (1997). The usefulness of the criteria-based content analysis technique in distinguish between truthful and fabricated allegations. *Psychology, Public Policy and Law*, *3*, 705–737.

Sandy, C., Rusconi, P., & Li, S. (2017). Can humans detect the authenticity of social media accounts? on the impact of verbal and non-verbal cues on credibility judgements of twitter profiles. In *2017 3rd IEEE International Conference on Cybernetics (CYBCONF)* (pp. 1–8). IEEE.

Sarna, G., & Bhatia, M. P. S. (2017). Content based approach to find the credibility of user in social networks: an application of cyberbullying. *International Journal of Machine Learning and Cybernetics*, *8*, 677–689.

Seth, A., Zhang, J., & Cohen, R. (2015). A personalized credibility model for recommending messages in social participatory media environments. *World Wide Web*, *18*, 111–137.

Sethi, R. J. (2017). Crowdsourcing the verification of fake news and alternative facts. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media* HT '17 (pp. 315–316). New York, NY, USA: ACM.

Sethi, R. J., & Rangaraju, R. (2018). Extinguishing the backfire effect: Using emotions in online social collaborative argumentation for fact checking. In *2018 IEEE International Conference on Web Services, ICWS 2018, San Francisco, CA, USA, July 2-7, 2018* (pp. 363–366). IEEE.

Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web* WWW '16 Companion (pp. 745–750). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, *9*, 1–9.

Shariff, S. M., Zhang, X., & Sanderson, M. (2017). On the credibility perception of news on twitter: Readers, topics and features. *Computers in Human Behavior*, *75*, 785–796.

Shiralkar, P., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2017). Finding streams in knowledge graphs to support fact checking. In V. Raghavan, S. Aluru, G. Karypis, L. Miele, & X. Wu (Eds.), *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017* (pp. 859–864). IEEE Computer Society.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, *19*, 22–36.

Silverman, C. (Visited January, 2019). Lies, damn lies and viral content.

Tavernisen, S. (2019). As fake news spreads lies, more readers shrug at the truth. *New York Times*, .

Thorne, J., & Vlachos, A. (2017). An extensible framework for verification of numerical claims. In A. Martins, & A. Peñas (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Software Demonstrations* (pp. 37–40). Association for Computational Linguistics.

Thorne, J., & Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3346–3359). Association for Computational Linguistics.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018a). Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 809–819). Association for Computational Linguistics.

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., & Mittal, A. (2018b). The fact extraction and verification (fever) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)* (pp. 1–9). Association for Computational Linguistics.

Tutek, M., Sekulic, I., Gombar, P., Paljak, I., Culinovic, F., Boltuzic, F., Karan, M., Alagić, D., & Šnajder, J. (2016). Takelab at semeval-2016 task 6: stance classification in tweets using a genetic algorithm based ensemble. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 464–468).

Vartapetiance, A., & Gillman, L. (2012). "i don't know where he is not": Does deception research yet offer a basis for deception detectives? In *Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection* EACL 2012 (pp. 5–14). Association for Computational Linguistics.

Viviani, M., & Pasi, G. (2017). A multi-criteria decision making approach for the assessment of information credibility in social media. In A. Petrosino, V. Loia, & W. Pedrycz (Eds.), *Fuzzy Logic and Soft Computing Applications* (pp. 197–207). Cham: Springer International Publishing.

Vlachos, A., & Riedel, S. (2014). Fact checking: Task definition and dataset construction. In C. Danescu-Niculescu-Mizil, J. Eisenstein, K. McKeown, & N. A. Smith (Eds.), *Proceedings of the Workshop on Language Technologies and Computational Social Science@ACL 2014, Baltimore, MD, USA, June 26, 2014* (pp. 18–22). Association for Computational Linguistics.

Vlachos, A., & Riedel, S. (2015). Identification and verification of simple claims about statistical properties. In L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, & Y. Marton (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015* (pp. 2596–2601). The Association for Computational Linguistics.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*, 1146–1151.

Vrij, A. (2000). *Detecting lies and deceit: the psychology of lying and implications for professional practice*. Wiley series in psychology of crime, policing and law. Wiley.

Vrij, A., Kneller, W., & Mann, S. (2000). The effort of informing liars about criteria-based content analysis on their ability to deceive cbca-raters. *Legal And Criminological psychology*, *5*, 57–70.

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 422–426). Association for Computational Linguistics.

Weerkamp, W., & de Maarten Rijke (2012). Credibility-inspired ranking for blog post retrieval. *Information Retrieval*, *15*, 243–277.

Weerkamp, W., & de Rijke, M. (2008). Credibility improves topical blog post retrieval. In *Proceedings of ACL-08* (pp. 923–931). Association for Computational Linguistics.

Wei, W., & Wan, X. (2017). Learning to identify ambiguous and misleading news headlines. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* IJCAI'17 (pp. 4172–4178). AAAI Press.

Wei, W., Zhang, X., Liu, X., Chen, W., & Wang, T. (2016). pkudblab at semeval-2016 task 6: A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 384–388).

Westerman, D., Spence, P. R., & Heide, B. V. D. (2014). Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication*, *19*, 171–183.

Wheeler, R., & Aitken, S. (2000). Multiple algorithms for fraud detection. In R. Ellis, M. Moulton, & F. Coenen (Eds.), *Applications and Innovations in Intelligent Systems VII* (pp. 219–231). London: Springer London.

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics* CICLing'05 (pp. 486–497). Springer-Verlag.

Wu, G., Greene, D., Smyth, B., & Cunningham, P. (2010). Distortion as a validation criterion in the identification of suspicious reviews. In *Proceedings of the First Workshop on Social Media Analytics* SOMA '10 (pp. 10–13). New York, NY, USA: Association for Computational Linguistics.

Wu, S., Liu, Q., Liu, Y., Wang, L., & Tan, T. (2016). Information credibility evaluation on social media. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* AAAI'16 (pp. 4403–4404). AAAI Press.

Wu, Y., Agarwal, P. K., Li, C., Yang, J., & Yu, C. (2014). Toward computational fact-checking. *Proc. VLDB Endow.*, *7*, 589–600.

Zarrella, G., & Marsh, A. (2016). Mitre at semeval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 458–463). Association for Computational Linguistics.

Zhou, F., Jin, J., Du, X., Zhang, B., & Yin, X. (2017). A calculation method for social network user credibility. In *2017 IEEE International Conference on Communications (ICC)* (pp. 1–6). IEEE.

Zhou, L., Burgoon, J., Nunamaker, J., & Twitchell, D. (2004a). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, *13*, 81–106.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker, J. F., Jr. (2004b). A comparison of classification methods for predicting deception in computer-mediated communication. *J. Manage. Inf. Syst.*, *20*, 139–166.

Zhou, L., & Zenebe, A. (2008). Representation and reasoning under uncertainty in deception detection: A neuro-fuzzy approach. *Trans. Fuz Sys.*, *16*, 442–454.

Zhou, L., & Zhang, D. (2008). Following linguistic footprints: Automatic deception detection in online communication. *Communications of the ACM*, *51*, 119–122.

Zhou, Y. (2017). Clickbait detection in tweets using self-attentive network. *CoRR*, *abs/1710.05364*, 1–5.

Zhou, Z., Guan, H., Bhat, M. M., & Hsu, J. (2019). Fake news detection via nlp is vulnerable to adversarial attacks. *arXiv:1901.09657*, .

CRediT autor statement

Estela Saquete: Conceptualization , Methodology, Investigation, Writing -Original Draft preparation, Reviewing and Editing, Supervision, Funding acquisition.

David Tomás: Conceptualization , Methodology, Investigation, Resources,Writing -Original Draft preparation, Reviewing and Editing, Visualization.

Paloma Moreda: Conceptualization, Methodology, Investigation, Resources, Writing -Original Draft preparation, Reviewing and Editing, Visualization.

Patricio Martínez-Barco: Conceptualization, Methodology, Investigation, Resources, Original Draft,Writing -Reviewing and Editing, Supervision, Project Administration.

Manuel Palomar: Conceptualization , Methodology, Supervision, Funding acquisition, Project Administration.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

2467

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: